Functional analysis of *E. coli* specific genes

Zubin Thacker

A thesis submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

The University of Edinburgh

2004

Acknowledgements

I would first like to thank the Darwin Trust Scholarship and in particular Prof. Ken Murray for their support, without which pursuing this PhD would have proved impossible. I am also indebted to various staff members of the ICMB and University who through their efforts have made my time in Edinburgh fruitful and enjoyable. I would like to extend my thanks to Sean McAteer, Garry Blakely, Willie Donachie and Andrew Coulson whose forthcoming advice and support have helped shape my Ph.D. I am especially thankful to Christophe and Millie for their supervision, guidance and friendship. I would also like to thank the many friends I have made both in and outside the department without whom the last four years would have been unbearable. My parents deserve special mention for their unquestioning and constant support. And finally I would like to thank Jen for understanding when I ranted, for her support, but most of all, for her love.

Contents

List of tables

List of figures.

To my grandmother (pakit mummy).

ABSTRACT

Bacteria belonging to the diverse gamma-proteobacterial family show genome sizes that vary from 640 Kb (*Buchnera* sp.) to 6.3 Mb (*Pseudomonas aeruginosa*). *E. coli* is an important member of the gamma-proteobacterial family and is closely related to pathogenic enteric organisms like *Shigella* and *Salmonella.* It occurs naturally in the colon of humans and other vertebrates and certain serotypes cause diseases of the enteric, pulmonary, nervous and urinary systems in humans. While genes shared between species that may or may not be involved in pathogenesis have received a great deal of interest, genes that are specific to a single species have largely been ignored. The basis of my research has been to identify and characterise the function of '*E. coli* specific' genes. *E. coli* specific genes were identified after comparing 33 gamma proteobacterial genomes at the Microbial Genome Database (MBGD). The database was also used to investigate the change in the number of *E. coli* specific genes since the completion of the first *E. coli* genome in 1997. Forty nine selected *E. coli* specific genes were deleted in 38 seperate deletion events on the genome of the model *E. coli* K-12 MG1655 using the modified pKO3 deletion procedure (Merlin et al. 2002). Gene expression levels and patterns in various phases of growth in LB broth were measured. Mutants and parent strains were tested for growth in a variety of conditions in agar based media. A mutant of the *yigE* ORF was found to be sensitive to high concentrations of nickel and cobalt in the growth medium. Mutation of the *E. coli* specific *htrC* gene showed no temperature sensitivity as reported in published literature and is shown here to play no part in the heat shock response of *E. coli.* The chromosomal position of the acetate utilisation gene *ackB* (Pascal et al. 1981) is experimentally demonstrated to be close to the known acetate kinase gene *ackA* at 50 minutes on the K-12 chromosome. Mutations in ORFs *yceP, ydeK* and *ygjMN* caused sensitivity to high levels of the basic dye crystal violet. This study shows that the group of genes specific to *E. coli* targetted here are expressed, show different patterns of expression and some are phenotypically functional. This study draws attention to a fraction of genes whose functional contribution to *E. coli* may have been underestimated due to their poor conservation in genomes other than *E. coli.*

Chapter 1: Introduction

1.1. Bacterial diversity and genome sequencing.

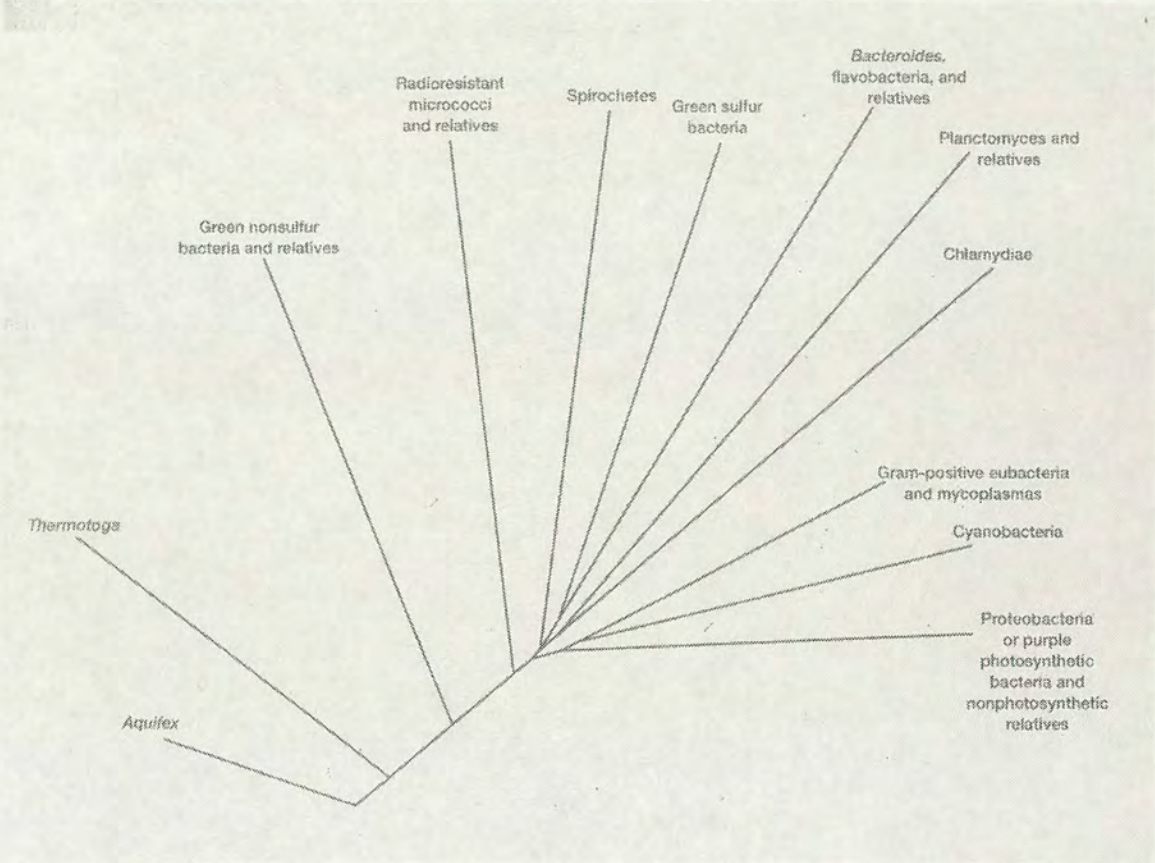
Microscopic life on earth began almost 4 billion years ago and has been evolving since. During this time microbial organisms have evolved to become the most versatile forms of life known. They can exist in extreme environments (geothermal vents, ice fissures, chemically and radioactively contaminated areas etc.), as harmless or beneficial commensals and obligate or virulent pathogens of animals and plants. Their wide diversity means that although invisible to the naked eye, they form an essential part of our ecosystems.

One of the first steps in understanding such a diverse life form is to classify or arrange them into groups of members on the basis of relatedness. Methods involving analysis of their physiology, morphology and ecology have been useful, to a certain degree, in categorising microbes into broad groups (prokaryotes and eukaryotes). The discovery by Carl Woese and colleagues in 1977, that ss 16S rRNA (small subunit ribosomal RNA) molecules could be used to analyse evolutionary distances between bacteria (and between other forms of life), has led to bacterial classification as we understand it today. 16S rRNA analysis divided life into three independently evolving branches, eubacteria, archaebacteria and eukaryota (Woese & Fox, 1977). Further analysis of ss rRNA has produced the phylogenetic tree of eubacterial organisms shown below.

Although not without controversy, 16S rRNA analysis has provided a glimpse into the possible evolutionary relationships that exists between microbes. However, it is the dawn of large scale, whole genome sequencing, that has made it possible to compare entire genomes and the evolutionary relationships that exist between genomes. The genome sequence of the microbe *H. influenzae* was the first completed and published genome for any living organism. Since that time 229 genomes have been fully

sequenced and there are currently a reported 525 microbial genomes being sequenced (http://www.genomesonline.org/).

Figure 1.1. Phylogeny of bacteria as suggested by 16s RNA analysis



The advantages of obtaining full genome sequences are many. They open a rich mining ground to search for industrially and environmentally applicable technologies. Understanding mechanisms that allow archaebacteria and eubacteria to survive in exotic environments could allow the development of new industrially applicable enzymes. Comparing full genome sequences of pathogens, identifying genetic islands associated with disease and understanding the genetic relationships that exist between hosts and pathogens holds the potential to revolutionise our understanding of pathogenesis and our ability to combat it.

1.1.1. Understanding genome evolution through sequence analysis.

One immediate advantage of having numerous bacterial genome sequences for analysis is the opportunity to understand the pathways and results of millions of years of microbial evolution. Since genome sequencing provides a global perspective of microbial genomes, it is possible to track large-scale lateral transfer and other mutational events that lead to genome evolution. The contribution of these events to genome evolution could not be appreciated earlier by only comparing single representative 16S rRNA sequences. An example is the genome of the eubacterial thermophile *Thermotoga maritima* whose sequence has revealed that 24% of its genes are most similar to archaeal genes (Nelson et al., 1999).

Full genome sequencing has also proven essential to understanding the effects of smaller lateral transfer events in niche adaptation and speciation. An example is the identification of genes coding for an archaeal like bacteriorhodopsin (archaeal halophilic membrane protein which functions as a light driven proton pump) in an uncultivated γ-proteobacterium (naturally occurring marine bacterioplankton) indicating new types of photoautotrophy in the sea (Beja, 2000). The *lac* operon and the ability of *E. coli* to use lactose as sole carbon source has been shown to have been gained by lateral transfer (Ochmann et al, 2000). Indeed full genome sequencing has provided evidence of distinct laterally transferred ribosomal RNA operons in the genome of a thermophilic actinomycete *Thermospora chromogena* (Yap et al., 1999).

Full genome sequences are, for the first time, revealing the true nature of diversity in microbes. It is highlighting how microbes are related to each other and yet unique due to the evolution of a unique composition of genes and sequences in each genome. Each genome has a certain fraction of genes that are essential for survival and are strongly conserved between species. However, evolution has also resulted in another fraction of

genes that are horizontally shared and conserved in very closely related species. The conservation patterns of these fractions of genes are similar for closely related organisms and therefore have encouraged the need to develop a species genome concept (Boucher et al, 2001; Lan & Reeves, 2000).

1.2. The Escherichia coli *group of organisms and its diversity:*

The group of organisms belonging to this species are Gram negative facultative anaerobic rods that occupy a wide variety of ecological niches. Members of this species are found as commensals of mammals, survive freely in the environment and also cause infections of the enteric, pulmonary, nervous and urinary systems in humans. The ability to survive in very diverse environments exhibited by members of this species and the evolutionary pressures that accompany niche adaptation have shaped the genomes of the various strains of *Escherichia coli*. Reference collections of environmental isolates of *E. coli* (ECOR strains) have genome sizes that range from 4.5 to 5.5 mega base pairs (Bergthorsson and Ochman, 1998). Pathogenic strains of *E. coli* also show a large amount of genome plasticity and diversity of gene content which mirror the variety of diseases of the enteric, urinary and nervous systems in humans that these strains cause (Rode et al, 1999).

Table 1.1. E. coli *strains and their genome sizes (Dobrint et al, 2003).*

Pathotype	Strain	Chromosomal size (Mb)
UPEC (Uropathogenic *E. coli*)	536	4.92
UPEC	536-21	4.69
UPEC	J96	5.11
UPEC	J96-M1	5.06
UPEC	764	5.07

UPEC	764-2	5.02
UPEC	P42	5.06
UPEC (ABU)	83972	4.88
MNEC (Meningitis associated *E. coli*)	IHE3034	4.84
Fecal isolate	F54	5.02
ETEC (Enterotoxigenic *E. coli*)	C9221a	4.79
EPEC (Enteropathogenic *E. coli*)	E2348/69	4.7
EIEC (Enteroinvasive *E. coli*)	EDL1284	4.68
EHEC (Enterohaemorrhagic *E. coli*)	4797/97	5.25
EHEC	5714/96	4.89
EHEC	1639/77	4.85
EHEC	SF493/89	4.87
EAEC (Enteroaggregative *E. coli*)	DPA065	4.69
Commensal isolates	MGS 6	4.71
	MGS 32	4.95
	MGS 73	4.99
	MGS 89	5.03
	MGS 104	5.03
	MGS 124	5.07
Laboratory strains	B	4.7
	MG1655	4.64

Today with the advent of full genome sequencing the genomes of five *E. coli* strains are available on public databases for download and analysis; two non-pathogenic laboratory strains K-12 MG1655 (Blattner et al, 1997) and W3110 (Itoh et al, 1999), two enteric pathogens 0157:H7 (Hayashi et al, 2001) and EDL933 (Perna et al, 2001) and the uropathogenic strain CFT073 (Welch et al, 2002). The nonpathogenic K-12 strain of *E. coli* is a model organism with a long history of laboratory research going back 50 years. Its experimental and biotechnological importance were the reasons it was one of the first

organisms chosen for full genome sequencing. The two 0157 strains are potent enteric pathogens of humans but asymptomatically colonise the gut of ruminant organisms (Naylor et al, 2003). The final sequenced strain of *E. coli* is CFT073, which is a potent pathogen causing neonatal meningitis and urinary tract infections in humans.

Comparing the genome of K-12 MG1655 to that of strains 0157 and CFT073 shows that the two pathogenic strains have larger genome sizes 5.5 and 5.3 megabases compared to the genome of strain K-12 which is 4.6 megabases in length. The larger genome sizes of the pathogenic strains are due to numerous regions on the chromosomes that have been horizontally acquired and are implicated in disease. Much of this DNA occurs in contiguous segments termed 'islands'. Islands specific to pathogenic strains and carrying pathogenicity determinants are termed pathogenicity islands (PAI). Islands show dissimilar gene content between the 0157 and CFT073 strains reflecting the different diseases caused by the two strains. All three strains however share a large proportion of the chromosome which forms the so called 'backbone'. Strains 0157 and CFT073 share 4.1 and 3.92 megabases of the backbone, which is essentially colinear, with the non-pathogenic K-12 MG1655 strain (Welch et al, 2002).

Full genome comparison of the K-12 MG1655 genome with the closely related pathogenic *Shigella flexneri 2a* genome has also shown a similar colinearlity of 3.9 Mb. The pathogenicity of the *Shigella* strain is attributed to numerous inversions, translocations and insertions which modify the chromosome and allow for expression of pathogenesis genes carried on a plasmid (Jin et al 2002). Development of pathogenesis therefore involves not only the acquisition of genes encoding proteins involved in pathogenesis but also in alterations to the chromosome that cause loss of function enabling pathogenesis. An example is the loss of gene function encoded by *cadA* which codes for protein lysine decarboxylase which converts lysine to cadaverine. Expression of *cadA* does not affect the invasive capability of *Shigella,* however it inhibits *Shigella* enterotoxin activity. Loss of this gene and its encoded protein is essential for pathogenesis caused by *Shigella* (Day et al, 2001).

Full genome comparisons show that although the K-12 strain is normally non-pathogenic, the right combination of gene gain and loss on the chromosome can lead to the evolution of diverse pathogens. The 'backbone' of shared genome sequence acts as a base to which 'modules' of pathogenesis may be added. The backbone is of interest because unlike pathogens such as *Yersinia pestis* and *Mycobactrium leprae* that show degenerate genomes (Chain et al, 2004; Cole et al, 2001) the *E. coli* pathogens show very little degeneration in terms of gene loss by insertions and inversions. This is perhaps due to the role played by genes on the 'backbone' in survival in the gut or the external transient environment.

The work carried out in this project was aimed at functionally analysing a fraction of genes or open reading frames that form part of the backbone shared by and specific to all sequenced *E. coli* genomes. Genes called 'specific' here are those that are found in all *E. coli* genomes but are absent from closely related gamma proteobacterial species. Functionally characterising these genes may reveal functions and pathways unique to the *E. coli* group of organisms that may provide new approaches to identifying and combating *E. coli* related illnesses.

1.3. Species specific or ORFan genes in E. coli.

When the complete genome sequence of *E. coli* K-12 was published in 1997 about 60% of the genes on the chromosome appeared to be 'specific' to *E. coli* as they were not found in the five other completed genomes available at the time. These genes were then described as "a subset of proteins specific to enterobacterial or *E. coli* processes as well as insertion elements and phage with host range restricted to *E. coli*" (Blattner et al, 1997). The large increase in genome sequence information since then has shown that rather than decreasing the number of genes that appear to be ORFans (specific to certain closely related strains or species and of unknown function) the

number of ORFan genes has been increasing. In a study where 60 fully sequenced microbial genomes were analysed there were 23,634 sequences found to be unique, although the percentage of ORFans as a fraction of all sequences has fallen to 14% (Siew and Fisher, 2003).

Since a significant number of full genome sequences have been available many studies have been devoted to understanding the functions of genes that bioinformatic tools identify as conserved in most or all genomes. Functional analysis of these genes has been of interest due to the hypothesis that widely conserved genes are likely to encode essential proteins. Although there is evidence now to support this hypothesis (Gerdes et al, 2003), it has been found that highly conserved genes are not always functionally essential under laboratory conditions (Arigoni et al, 1998 & Frieberg et al, 2001)

Species specific or ORFan genes on the other hand have received little attention until recently. Their presence in microbial genomes in the past had been attributed to sparse sampling and limited genomic data. However, since the number of ORFan genes has continued to grow, other explanations are necessary in order to understand the origin and function of this class of genes.

There have been increasing numbers of studies demonstrating that ORFan genes are indeed expressed and functional. One French group has been particularly successful in compiling gene expression data, crystal structures and functional analyses of ORFan genes in *E. coli*. In a study using reverse transcriptase polymerase chain reaction to detect messenger RNA, 19 out of 25 ORFan genes selected were shown to be expressed (Alimi et al, 2000). This led the researchers to crystallise one of the ORFan proteins, encoded by the annotated gene b0220, that exhibited a high level of expression during the exponential and stationary phases of growth (Abergel et al, 2000). Another ORFan

protein YkfE was functionally annotated as an inhibitor of C type lysozyme after a serendipitous discovery during a routine protein crystallization procedure (Monchois et al 2001). Although these appear to be small steps in functionally characterising a growing population of ORFan genes each step has revealed previously unknown or unexpected patterns of gene expression, protein folding or protein function.

The growing attention the ORFan genes are receiving is reflected by the web based databases cataloging the global distribution of ORFan genes. The ORFan database (http://www.cs.bgu.ac.il/~nomsiew/ORFans/#SinO) catalogs ORFan genes (single, paralogous and orthologous) in 84 bacterial genomes available at the NCBI database. The database shows how the number of ORFans has changed in each genome as the database has grown to include 60 and later 84 genomes. The web page shows only the results of precomputed ORFan percentages in sequenced microbial genomes according to parameters described. It does not allow the user to change parameters or reanalyse genomic sequences.

The other relevant web based database that I have consulted is called Neurogadgets (www.neurogadgets.com/bws.php). Unlike the ORFan database mentioned above this database is not dedicated to ORFans but is a web front to bioinformatic analyses of genomic sequences in the Neurogadgets database. One of the bioinformatic analyses offered is the identification of ORFans in the various genomes on the database. This database allows the user to define the cut-off parameters used to differentiate real matches from random matches. It also allows user selection of genomes to compare against a query genome. It is therefore more interactive and useful than the ORFan database.

Both the above internet resources utilise bioinformatic tools to identify and describe ORFan genes. To functionally characterize these genes and understand their contribution to the bacterial cell is the job of molecular biologists. The ORFan genes represent a unique problem to molecular biologists and require novel methods of investigation. The

more established method of functionally characterising genes of unknown function, by looking for matches to orthologs or homologs with known function, obviously cannot be applied to ORFan sequences. The databases mentioned above are a starting point in investigating the function and evolution of ORFan genes since they provide the only possible way to analyse the ever growing sequence databases to identify ORFan genes.

Of interest to an evolutionary and/or molecular biologist are questions regarding the origin and rates of divergence of ORFan sequences. Do ORFan sequences represent conditionally essential genes necessary for niche adaptation, or are they novel regulatory regions which are not expressed? Do ORFan sequences represent once, but now no longer, functional DNA, or do they encode proteins with potentially novel functions or indeed, whole pathways which would help to better understand *E. coli* or any other organism of interest? And since sets of ORFan sequences remain restricted to a small group of organisms or groups are they involved in species determination? These are some of the questions which this study was undertaken to answer.

1.4. E. coli K-12 MG1655; a model genetic organism.

Experimental work carried out in this study was done on the model *E. coli* K-12 strain MG1655. K-12 strains are non-pathogenic and have more than 50 years of history as a preferred model for biochemical, molecular biology and biotechnology research. Genetic manipulations in the laboratory have led to the loss of F plasmid, and of bacteriophage lambda, creation of the pyrimidine starvation phenotype, disruption of the isoleucine valine biosynthesis pathway and loss of O-antigen synthesis in the LPS (Blattner et al, 1997).

Full genome sequencing of strain K-12 MG1655 completed in 1997 showed that nearly 40% of predicted open reading frames were not associated with a known function (Blattner et al, 1997). A more recent update of the K-12 chromosome shows the genome

having 4,401 genes specifying 116 RNAs and 4285 proteins. Since some protein coding regions are compound and encode multimodular proteins the 4,401 genes encode 4,616 modules. Of the modules nearly 49% have no known or experimentally verified function with 29.5% of modules having an imputed function only and 19.5% of modules lack even a predicted function (Serres et al 2000). The functional breakdown of the *E. coli* chromosome as reported by Serres et al (2001) is shown in table 1.2.

Table 1.2. Functional break down of E. coli *genes (Serres et al, 2001).*

Category	Experimental	Putative	Total
Enzymes	990	550	1540
Transport proteins	310	298	608
Regulators	213	151	364
Membrane	47	132	179
Factors	109	33	142
Structure	90	35	125
Carriers	35	25	60
External Origin	282		282
Phenotype	98		98
RNA	116		116
Leaders	12		12
ORFs	886		886
Total	3188	1224	4412

The large fraction of genes on the chromosome of *E. coli* K-12 MG1655 that remain functionally unknown need to be studied and their contribution to the cell understood. *E. coli* strain K-12 is one of the best biological candidates whose physiological and genetic make-up is close to being fully understood. Discussed below are approaches in use for functional analysis of the *E. coli* chromosome.

1.5. Functional genetics of the E. coli chromosome.

The rate of full genome sequencing and annotation has created a bottleneck in the field of functional genetics because genes are now sequenced faster than they can be functionally characterised. Advancements in functional analysis of genomes now allow genes or entire genomes to be compared to other genomes across a range of sequenced genomes. Comparison tools are evolving as are protocols to simulate entire genomes on the computer. Advances in molecular and protein sciences have provided scientists with high-throughput and sensitive methods to study gene transcript levels and identify proteins expressed in different conditions. Processes of gene deletion and mutant testing have also been scaled up to make this previously laborious task a reliable and reproducible approach. Discussed below are some advances in functional analysis of the *E. coli* chromosome.

1.5.1. Bioinformatics.

Computers, computer programs and algorithms have been crucial to the advancement of genetic research. Managing, storing, analysing, annotating and modeling full genome contents of the numerous sequenced biological species would be impossible without the use of computers and related programs. Discussed below are a range of available databases and programs useful in functional genetic research of the *E. coli* and indeed other genomes.

1.5.2. Databases for *E. coli* gene and genome sequence retrieval and analysis.

Computers have proven essential in cataloging, storing and accessing the vast amounts of data generated in genome sequencing projects along with related sequence annotation and literature. Many databases associated with the genome sequence,

proteome and metabolome of *E. coli* are available over the internet. I have summarised some of the most significant resources below.

- *Colibase (http://colibase.bham.ac.uk/):* Like the Colibri database, Colibase provides free internet based services for sequence query, access, analysis and downloads. However unlike the Colibri database, Colibase, being a more recent development, also stores full genome sequence and annotation of multiple *E. coli, Salmonella* and *Shigella* genomes. The website also has links to pages on the Swiss-prot (http://ca.expasy.org/) and Genbank databases (http://www.ncbi.nlm.nih.gov/).

- *Colibri (http://genolist.pasteur.fr/Colibri/):* This was one of the first online databases created for remotely accessing and analysing the *E. coli* K-12 MG1655 chromosome. The database has programs for viewing and downloading nucleotide and protein sequences and also holds additional information on each gene such as its codon adaptation index (CAI), calculated isoelectric point (pI), codon usage and links to corresponding pages on the Swiss-prot (http://ca.expasy.org/) website. Other services include nucleotide and protein Blast analysis of query sequences against the K-12 chromosome on the database.

- CGSC database (http://cgsc.biology.yale.edu/): The CGSC Database of *E.coli* genetic information includes genotypes and reference information for the strains in the CGSC collection, gene names, properties, and linkage map, gene product information, and information on specific mutations.

- *E. coli* cell envelope protein data collection, ECCE (http://www.cf.ac.uk/biosi/staff/ehrmann/tools/ecce/ecce.htm): Collection of data on the functional classification of cell envelope proteins.

- *E. coli* index (http://ecoli.bham.ac.uk/): Comprehensive guide to information relating to *E. coli*; provides links to useful databases for researchers and the general public.

- EcoCyc (http://ecocyc.org/): is a scientific database for the bacterium *Escherichia coli* K12 MG1655. The EcoCyc project performs literature-based curation of the entire genome, and of transcriptional regulation, transporters, and metabolic pathways.

- *E. coli* genome project (http://www.genome.wisc.edu/): Web resource maintained by the same research group that sequenced the K-12 genome. Now lists annotations, updates and functional genomic data obtained from deletion constructions and microarray expression experiments.

- Ecogene (http://bmb.med.miami.edu/EcoGene/EcoWeb/): Ecogene is described as "A collection of information about the genes, proteins, and intergenic regions of the *E. coli* K-12 genome and proteome accumulated during years of sequence analysis and literature surveys by Kenn Rudd and his collaborators: Mary Berlyn of the *E. coli* Genetic Stock Center, Amos Bairoch of SWISS-PROT, and Antoine Danchin and Ivan Mozser of Colibri."

- *E. coli* genome and proteome database, GenprotEC (http://genprotec.mbl.edu/): Database dedicated to the functions encoded by genes of *E. coli* K-12 MG1655. Allows gene searches, provides overviews of the chromosome and provides classification of proteins into functional modules or groups.

- *E. coli* protease database, EPD (http://www.cf.ac.uk/biosci/staff/ehrmann/tools/proteases/index.html)
  Useful resource for all known and predicted proteases of *E. coli.*

- Genobase (http://ecoli.aist-nara.ac.jp/): Functional genomic analysis of *E. coli* in Japan. This web resource collates data and provides updates of the progress of this Japanese consortium formed to analyze the *E. coli* genome.

- Microbial genome database for comparative analysis (MBGD, http://mbgd.genome.ad.jp/): Designed for comparative analysis of microbial genomes on the database. This web resource has been used extensively in this study and is described in more detail in Chapter 3.

- Oklahoma University *E. coli* gene expression database (http://chase.ou.edu/macro/): Web based resource for the storage and retrieval of microarray mediated gene expression data. Links to supplementary microarray data from published literature.

- Profiling the *E. coli* chromosome: PEC (www.shigen.nig.ac.jp/ecoli/pec/index.jsp): Services include downloading and viewing *E. coli* DNA sequences. This web based resource also has grouped 4,409 *E. coli* genes into three categories, essential (252), non-essential (2,368) and unknown (1,789), based on published literature.

- Regulon DB database (http://www.cifn.unam.mx/Computational_Genomics/regulondb/): A database of transcriptional regulation and operon organisation of the *E. coli* chromosome.

1.5.3. Computer programs for gene and genome sequence analysis.

Along with storing and cataloging sequence and sequence information computers have also proven essential in comparative analysis, modeling and prediction of gene function. One of the programs most commonly used by molecular biologists is the Basic Local Alignment Search Tool (BLAST) described by Altschul et al (1990). It is essentially a computer algorithm used to compare a query nucleotide or protein sequence to one or many subject sequences within a defined database. Alignments between the query and subject sequences are then scored based on the degree of similarity between the two. This approach is useful in functional genetics since matches of the query sequences of unknown function to subjects with known functions suggests shared or similar functions.

The program can also be used to investigate degrees of gene conservation in different genomes. The BLAST program has been used countless times in gene and genome annotation and for selecting targets for functional analysis.

The BLAST program has been expanded to compare entire genomes (either one-to-one or one-to-many). These comparisons are useful in understanding how processes of gene deletion, duplication, insertion or inversion shape microbial genomes during evolution. The TIGR-CMR website has online programs which, using BLASTn, can compare a selected genome to all other microbial genomes on the database (http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl). Artemis and its associated program the Artemis Comparison Tool (ACT) available at the Sanger Center website (http://www.sanger.ac.uk/) also use BLASTn to compare two or more bacterial genomes. The programs then generate a cartoon of results of the BLASTn search which helps to visualise the differences in the two compared genomes.

Advances to the BLAST search tool can be used to search for matches in all possible reading frames of a sequence (BLASTx) or for short strongly conserved matches (gapped BLAST). Alternatives to BLAST such as the program MPsrch (www.ebi.ac.uk/MPsrch) use advanced Smith-Waterman algorithms that claim to be faster and more accurate. However BLAST remains one of the most commonly used programs by geneticists and molecular biologists.

Newer methodologies using computers to understand, predict and model bacterial genetic systems have been evolving. Simulating the entire *E. coli* cell and its metabolic and genetic pathways is one such project in development which uses chemical and mathematical equations collectively called flux-balance analysis. This modeling approach is based on the concept that cells obey the laws of physics and chemistry which include conservation of mass, energy and redox potential. Such laws constrain cellular behavior and a system whose boundaries of cellular capabilities are known can be modeled (Kauffman et al, 2003). This modeling approach has been used to

demonstrate the robustness of *E. coli* to changes in individual enzyme or pathway activities. Although the system is some way from being perfected since extensive gaps in the knowledge of metabolic and genetic pathways remain, the system is reported to be capable of being expanded to include new data as it becomes available (Edwards and Palsson, 2000).

1.5.4. Advances in experimental genetic and proteomic methodologies.

Predictions of gene function obtained using computer models and comparisons need to be tested. It is the information derived from molecular genetic and proteomic research that ultimately increases our understanding of gene, pathway or protein functions. Discussed below are some significant advances in experimental genetic and proteomic methodologies that are being used to test and define gene function.

1.5.5. DNA microarray technologies:

Full genome sequencing, and advances in robotics, computing and molecular methodologies have enabled DNA microarray technologies. Every known and predicted gene on a genome can be represented on a slide which can then be used to measure the mRNA transcript levels of every gene in a variety of growth or test conditions. The great benefit of this technology is that it provides a global view of gene transcript levels in response to environmental or test conditions. This whole genome information can be used to construct a transcript profile of genes which corresponds to a functional role of the respective proteins in specific test conditions.

High costs of equipment and running materials, standardising operational variations, managing and analysing the large datasets produced with effective statistical tools to differentiate between 'real' expression changes and those that emerge from experimental

errors remain as drawbacks of this system. However such drawbacks have not prevented the adoption of this very powerful method (Dharmadi & Gonzalez, 2004).

DNA microarrays have been used to measure the global response of *E. coli* to a variety of different conditions A search on the NCBI-Pubmed database for the keywords 'microarray' and '*Escherichia coli*' shows a large set of published literature using this technology. Many of the published works are devoted to measuring global gene expression in response to physical or other conditions (temperature, pressure, antibiotics, xenobiotics etc.) or genetic conditions such as effects of deletion or overexpression of known gene regulators. Some published microarray based literature on *E. coli* is shown in table 1.3. below.

Table 1.3. Microarray based research on the E. coli *chromosome..*

Gene expression studies of environmental stimuli.	**Gene expression in different genetic conditions.**
Richmond et al, 1999: Effect of heat shock and IPTG treatment.	Khodursky et al, 2000: Genetic effects of tryptophan metabolism.
Tao et al, 1999: Growth on LB and minimal media with glucose.	Oh and Liao, 2000: Effects of protein overproduction.
Wei et al, 2001: Growth in exponential and transitional phases, induction with IPTG.	Arfin et al, 2000: Effects of Integration Host Factor (IHF).
Zheng et al, 2001: Response to H_2O_2.	Zimmer et al, 2000: NtrC regulated genes.
Oh et al, 2002: Acetate responsive genes.	Wei et al, 2001: Impact of *sdiA* amplification.
Bernstein et al, 2002: Analysis of mRNA decay and abundance.	Hommais et al, 2001: H-NS monitored gene expression.
Tani et al, 2002: Adaptation to famine.	DeLisa et al 2001: Genes controlled by autoinducer 2-stimulated quorum sensing

	in *Escherichia coli.*
Phadtare et al, 2002: Response to 4, 5-dihydroxy-2-cyclopenten-one.	Reitzer and Schneider, 2001: Sigma 54 dependent genes.
Rozen et al, 2002: Response to sea-water.	Martin and Rosner, 2002: The MarA/SoxC/Rob regulon.
Cheung et al, 2003: Osmotic stress response and supercoiling.	Lehnen et al, 2002: LnhA – Transcriptional regulator of flagella, motility and chemotaxis.
Polen et al, 2003: Long term adaptation to acetate and propionate.	Inoue et al, 2002: Growth inhibition by acetate of E. coli expressing Era-dE
Schembri et al, 2003: Gene expression in biofilms.	Hung et al, 2002: Effects of LRP.
Masuda and Church, 2003: Acid resistance genes in E. coli.	Oshima et al, 2002: DAM controlled gene expression.
Salmon et al, 2003: Oxygen availability and FNR.	Schembri et al, 2002: Effects of *fim* mutations in E. coli.
Minagawa et al, 2003: Mg^{2+} stimulon in *E. coli.*	Oshima et al, 2002: Transcriptome analysis of all two component systems.
Ren et al, 2004: Gene expression in *E. coli* biofilms.	Masuda and Church, 2002: Response to EvgA.
Hua et al, 2004: Response to growth limiting nutrients in chemostat cultures.	Lobner-Olesen et al, 2003: Role of SeqA and Dam in *Escherichia coli* gene expression
Brokx et al, 2004: Genome-wide analysis of lipoprotein expression.	Eguchi et al, 2003: Genes regulated by EvgAS two component system.
Polen and Wendisch 2004, Genomewide expression analysis in amino acid-producing *E. coli.*	Liu and De Wulf 2004: the ArcA-P modulon.

Dahan et al, 2004: Transcriptome of enterohemorrhagic Escherichia coli O157 adhering to eukaryotic plasma membranes.	Manna et al, 2004: Microarray analysis of transposition targets in Escherichia coli.
Ishii et al, 2004: Effect of hydrostatic pressure.	Kabir et al, 2004: Effects of mutations in the rpoS gene on cell viability and global gene expression under nitrogen starvation in Escherichia coli.
	Hagiwara et al, 2004: Expression of BasS-R two component system controlled genes.

Microarray technologies are also being used for diverse purposes in *E. coli* such as discriminate between pathogenic and non-pathogenic *E. coli* strains (Wu et al, 2003), investigate diversity of pathogenic strains of *E. coli* and *Shigella* (Fukiya et al, 2004), identify single nucleotide polymorphisms in TEM-beta lactamases in *E. coli* (Grimm et al, 2004), find the correlation between gene expression and codon usage bias (dos Reis et al, 2003) and to identify conditionally essential genes (Tang et al, 2004).

1.5.6. Advances in Proteomics:

Proteomic research is seen as the 'next step' in the quest to completely understand a cell and the function of its constituents. Knowing transcript levels and regulatory patterns of genes in various conditions does not provide any information on the function of the genes in question. Knowing levels of transcript also provides no information on the post-transciptional modification or amount of the protein in the cell. Predicted genes on genomes remain hypothetical until their corresponding proteins are

identified. Finally it is the proteins within cells that perform biological functions and understanding protein function is critical to understanding the cell.

Advances in genome sequencing and bioinformatics have led to large advances in the field of proteomics. Protocols for protein identification now employ combinations of different techniques such as SDS-polyacrylamide gel electrophoresis (PAGE), mass spectrometry together with full genome sequence information to identify a large proportion of proteins within *E. coli*.

Advances in protein isolation and enrichment procedures have enabled characterisation of low abundance proteins in *E. coli* which were previously considered to be a major challenge in the complete proteomic characterisation of the cell (Fountoulakis et al, 1999). Improvements in sample preparation, electrophoresis and mass spectrometry (matrix assisted laser desorption ionisation – time of flight – MALDI-TOF) have been applied to studing the cell envelope and have resulted in one of the largest databases of membrane proteins of *E. coli* (Fountoulakis and Gasser, 2003). Emerging technologies include protein arrays which could potentially be used to study protein-protein or protein-DNA interactions and protein expression profiling (Macbeath & Schreiber, 2000; Zhu et al, 2001, 2003)

1.5.7. Advances in gene interruption and deletion technologies.

Prior to full genome sequence being available, chemical or physical mutagens were, and are still, used to randomly induce mutations in the *E. coli* chromosome. Mutagens act on DNA strands to produce insertions, point mutations, frame shifts or deletions of bases thus disturbing gene function in many ways. Some common mutagens used were ultraviolet radiation, nucleotide base analogs and transposons (Vinopal et al, 1987) and their modes of action are summarised in table 1.4 below.

Table 1.4. Commonly used mutagens and their effects on DNA (Parekh S, 2004).

Mutagen	Mutation induced	Impact on DNA	Relative effect
Radiation			
Ionising radiation			
1. X-rays, gamma rays	Single or double stranded breakage of DNA	Deletions, structural changes	High
Short wavelengths			
2. Ultraviolet rays	Pyrimidine dimerisation and cross-links in DNA	Transversion, deletion, frameshift, GC → AT transitions	Medium
Chemicals			
Base analogs			
3. 5-Chlorouracil, 5-bromouracil	Faulty base pairing	AT → GC, GC → AT transition	Low
4. 2-Aminopurine	Errors in DNA replication	-	Low
Deaminating agents			
5. Hydroxylamine (NH_2OH)	Deamination of cytosine	GC → transition	Low
6. Nitrous acid (HNO_2)	Deamination of A, C and G	Bidirectional translation, deletion, AT→GC and/or GC→AT transition	Medium
Alkylating agents			
7. NTG (N-methyl-N'-nitro-N-nitrosoguanidine)	Methylation, high pH	GC→AT transition	High

8. EMS (Ethyl methanesulphonate)	Alkylation of C and A	GC→AT transition	High
9. Mustards, di-(2-chloroethyl) sulphide	Alkylation of C and A	GC→AT transition	High
Intercalating agents			
10. Ethidium bromide, acridine dyes	Intercalation between two base pairs	Frameshift, loss of plasmids, microdeletion	Low
Biological			
11. Phage, plasmid, DNA transposons	Base substitution, breakage	Deletion, duplication, insertion	High

Since these mutagens mutagenise at random, connecting phenotypes to mutant genotypes was and is a challenge. Knowing full genome sequence has encouraged the development of tools for targeted mutation of genes to make the task of connecting phenotypes to genotypes more comprehensive. Each completely sequenced genome shows a significant number of predicted genes that have no function or phenotypes associated with them. Microarray and proteomic research can suggest conditions under which the genes or proteins are expressed however these suggestions of gene function also need to be tested at the genetic level. One of the most effective ways to test genes of unknown function is to delete/interrupt them and to test resulting mutants for phenotypic traits. Advances in random and targeted mutagenesis of genes in *E. coli* have made this a reliable and powerful approach to investigating gene function. Discussed below are some of these advances and a description of the gene deletion method used in this study.

1.5.8. Non targeted mutagenesis.

The ability of transposons to integrate randomly into double stranded DNA have made them an important tool in unselectively interrupting gene function. Transposon

mutagenesis of entire genomes has been applied to identify genes that are essential in a diverse range of organisms such as *Mycoplasma pneumoniae* and *Mycoplasma genitalium* (Hutchison et al, 1999), *Mycobacterium tuberculosis* (Sasetti et al, 2001) and *Saccharomyces cereviseae* (Ross-Macdonald et al, 1999).

Full genome transposon mutagenesis has been applied to the *E. coli* MG1655 genome to determine the number of essential genes. Of the 4,291 known and predicted genes on the chromosome, 3,746 genes were interrupted and 620 were found to be essential for logarithmic growth in aerobic conditions in enriched LB medium. Another finding of this study was a reportedly high correlation between essentiality of gene function and degree of conservation across species, in other words genes with high degrees of conservation were more likely to be essential (Gerdes et al 2003).

While this method has the advantage of being a high throughput system there are some disadvantages as well. Targeting genes below a certain length becomes difficult using transposon mutagenesis, as is the problem of completion since not all genes will be uniformly interrupted. Transposons are also more likely to integrate in certain regions of the chromosome (such as origins of replication) termed 'hotspots' than others. This may be due to higher target copy number at the origin of replication in an actively dividing bacterial population (Gerdes et al, 2003). However, despite these drawbacks transposon mutagenesis remains a very popular and powerful approach to functionally analyse entire genomes.

1.5.9. Targeted gene deletion.

The alternative to full genome transposon mutagenesis is to individually target every known or predicted gene in the chromosome of an organism. This approach although slower and more labour intensive has been used to study gene function in *Saccharomyces cereviseae* (Giaever et al, 2002, Winzeler et al, 2003), *C. elegans* (Kim, 2001) and *Bacillus subtilis* (Kobayashi et al, 2003). There are two different consortia

attempting to systematically mutagenise all known and predicted genes in the chromosome of *E. coli* K-12 MG1655. One group based at the University of Wisconsin headed by Dr. Frederick Blattner has reportedly deleted 2001 of 4,291 ORFs in strain MG1655 (as of 1.10.04: http://www.genome.wisc.edu/functional/tnmutagenesis.htm). This number includes 7 essential genes deleted using a conditional lethal amber mutation (Herring & Blattner, 2004). Another large collaboration of research groups based in Japan has also undertaken the task of understanding the role of every gene in the MG1655 chromosome by cloning, overexpressing, deleting genes and identifying the corresponding protein (Mori et al, 2000). An update of the progress of the Japanese consortium is pending.

Tools and methods for deleting selected genes on the chromosome of *E. coli* have made the methodical deletion of genes a reliable and reproducible task. Deletion of genes on the chromosome typically involves homologous recombination of a deletion construct created in-vitro using PCR and recombinant DNA technologies. The deletion construct is introduced into the cell as a circular or linear piece of double stranded DNA. Homologous recombination between the construct and its target results in the replacement of the target locus with the in-vitro deletion construct. Some commonly used and established methods of deletion are described below.

1.5.9.1. Gene deletion using linear double stranded DNA:

Deletion of genes using linear double stranded DNA generated in PCR reactions bearing 35-50 base pairs (bp) of homology with DNA regions flanking target genes has been possible in organisms such as *Saccharomyces cereviseae* (Szent-Gyorgyi et al, 1998) and *Candida albicans* (Wilson et al, 1999). This approach is attractive since it does not require any intermediate cloning steps involving DNA restriction and ligation.

However the method has been limited to deleting genes in organisms that have highly efficient and accurate mitotic recombination mechanisms.

Deleting genes in *E. coli* using linear double stranded DNA has been shown to work in wild-type (Jasmin and Schimmel, 1984) and in *recD* mutants (Russel et al, 1989). Recombination of linear DNA has also been possible in *recBC* and *sbcB* mutants of *E. coli* (Winans et al, 1985). Two separate studies have recently shown that it is possible to delete genes using linear PCR products that bear between 35-50 bp of homology to DNA flanking the target. Both approaches use the recombination genes *exo, bet* and *gam* encoded by the bacteriophage lambda. Each of the three genes play a role in the deletion process. *Gam* inhibits the exonuclease activity of RecBCD while *exo* degrades DNA at the point of double strand breaks in the 5'-3' direction. *Bet* binds to the remaining 3' end single stranded tail and protects and prepares it for homologous strand invasion (Yu et al, 2000, Datsenko and Wanner, 2000).

This procedure is faster than techniques that use cloned deletions on circular double stranded DNA and has been used to investigate gene function by many investigators (Skovran et al, 2004; Gong et al, 2003). However, problems include low efficiency of recombination which is seen in 0.1% of surviving cells from a standard electroporation. Low recombination frequencies also mean that steps during electroporation need to be optimised to maximise the chances of recombination. Deletion of essential genes also proves difficult using this method and can result in one of two possibilities. When deleting genes of unknown function a lack of recombinants could ambiguously mean that either the conditions were suboptimal or that the gene is essential. Alternatively recombinants carrying both the wild-type and mutant loci (Ellis et al, 2000) can also be obtained erroneously suggesting that the gene has been deleted and is not essential. It is for this reason that the method used in this study employs recombination between the chromosome and circular double stranded DNA described below.

1.5.9.2. Gene deletion using circular double stranded DNA.

Circular double stranded DNA molecules, namely conditionally replicating plasmids that are not targets of exonuclease activity, are the alternative vehicles for carrying deletion constructs into *E. coli.* In-vitro deletion constructs, which bear between 400-500 bp of homology to DNA flanking the target gene, are cloned onto conditionally replicating plasmids. The deletion vector is then transformed into the cell where integrants into the chromosome following spontaneous homologous recombination are selected (Hamilton et al, 1989). Several variations of this technique have been published and used to generate deletions, some of which are described below.

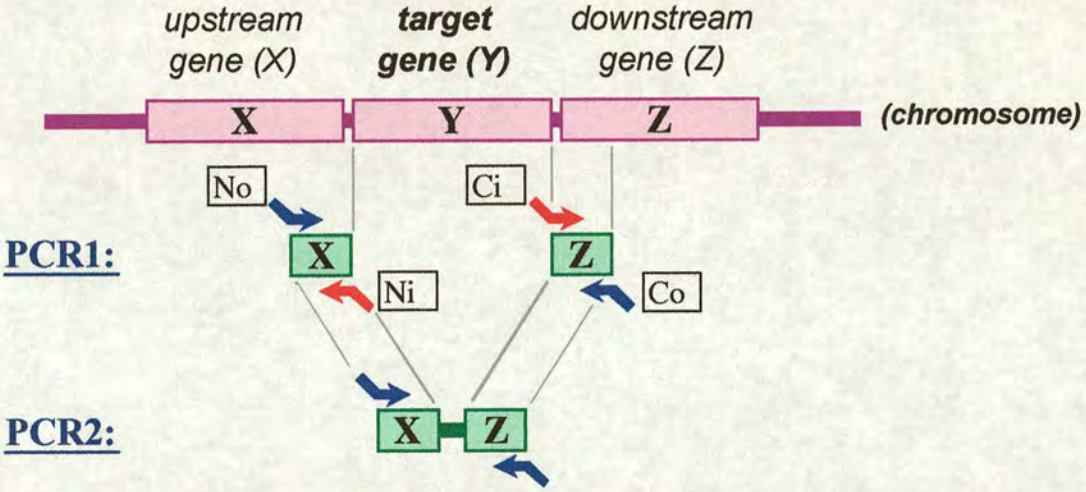
One method is to generate markerless deletions by cloning in-vitro constructs onto a suicide plasmid that carries a recognition site for the meganuclease I-SceI. Once the plasmid is inserted into the chromosome introducing a unique double strand break (DSB) into the chromosome at the I-SceI site stimulates the resolution process. The DSB enhances frequency of resolution and also selects for the resolved product (Posfai et al, 1999). A variation of this technique, called gene gorging, which uses the lambda *red* recombination system for gene replacements is reportedly efficient enough (1-15%) for direct PCR identification of recombinant progeny (Herring et al, 2003).

1.5.9.3. Gene deletion protocol described by Merlin et al, 2002.

The deletion method used in this study is a modified pKO3 based gene replacement protocol first described by (Link et al, 1997) and later improved upon by Merlin et al (2002). The deletion process begins by first creating an in-frame deletion of the gene of interest *in vitro*. Approximately 450-500 bp of regions flanking the gene of interest are amplified in PCR reactions using pairs of primers No-Ni and Ci-Co. N (N-terminal) and C (C-terminal) primers anneal upstream and downstream of the target gene, while "i" (inside) and "o" (outside) indicate whether the priming site is closer to or further from the target gene.

The inside primers are designed to leave the upstream and downstream ends of the target gene intact to ensure translational signals are left undisturbed. Ni and Ci primers also contain 24-nucleotide-long 5' tails with complementary sequences. The inside primers contain an internal *Not*I restriction site which is used to clone in a reporter cassette. Outside primers contain 11 nucleotide tails which provide convenient restriction sites to clone PCR products into the deletion vector pTOF24. Products of the 1st PCR reactions (1-2 μl) are used as substrates for the 2nd PCR reactions where the two arms are fused using primer pairs No and Co. This is illustrated in figure 1.2 below.

Figure 1.2. PCR reactions for in-vitro deletion construction.



A reporter cassette is cloned into the crossover product next using the internal *Not*I restriction sites. The reporter cassettes used in this study are FLK2 and FLKP2 as constructed by Merlin et al (2002). Reporter cassette FLK2 has the *lacZ* (β galactosidase) and *aph* (kanamycin resistance), genes flanked by FRT (FLP recombinase target) sites. Reporter cassette FLKP2 is identical to FLK2 but has an additional *plac* promoter between *aph* and FLP recombinase recognition sites (hatched boxes). The *plac* promoter was used to minimise any loss of expression of genes downstream of the deleted target. Shown below (Figure 1.3.) is a cartoon of the reporter cassettes used in this study. The deletion vector pTOF24 is a derivative of the vector pKO3 described by

Link et al (1997). Figure 1.4. shows the cloning strategy used to create the final deletion vectors used in this study.

Figure 1.3. Reporter cassettes FLK2 and FLKP2

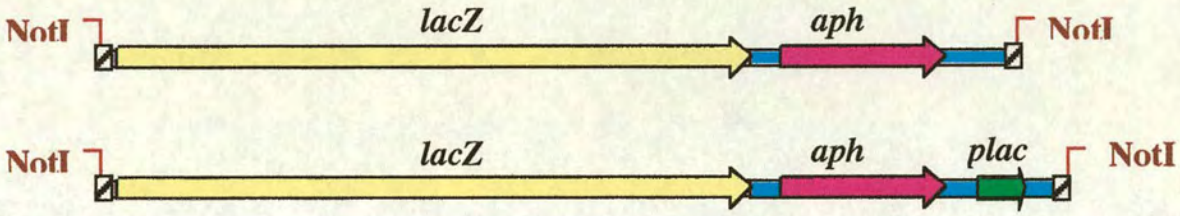
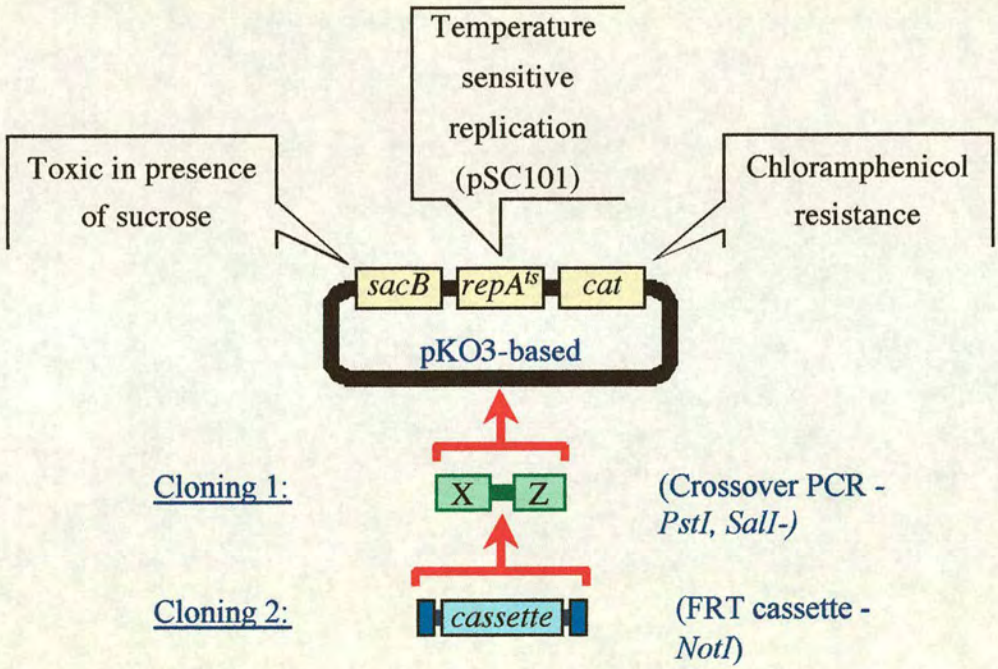


Figure 1.4. Cloning strategy for the construction of gene deletions in vitro.

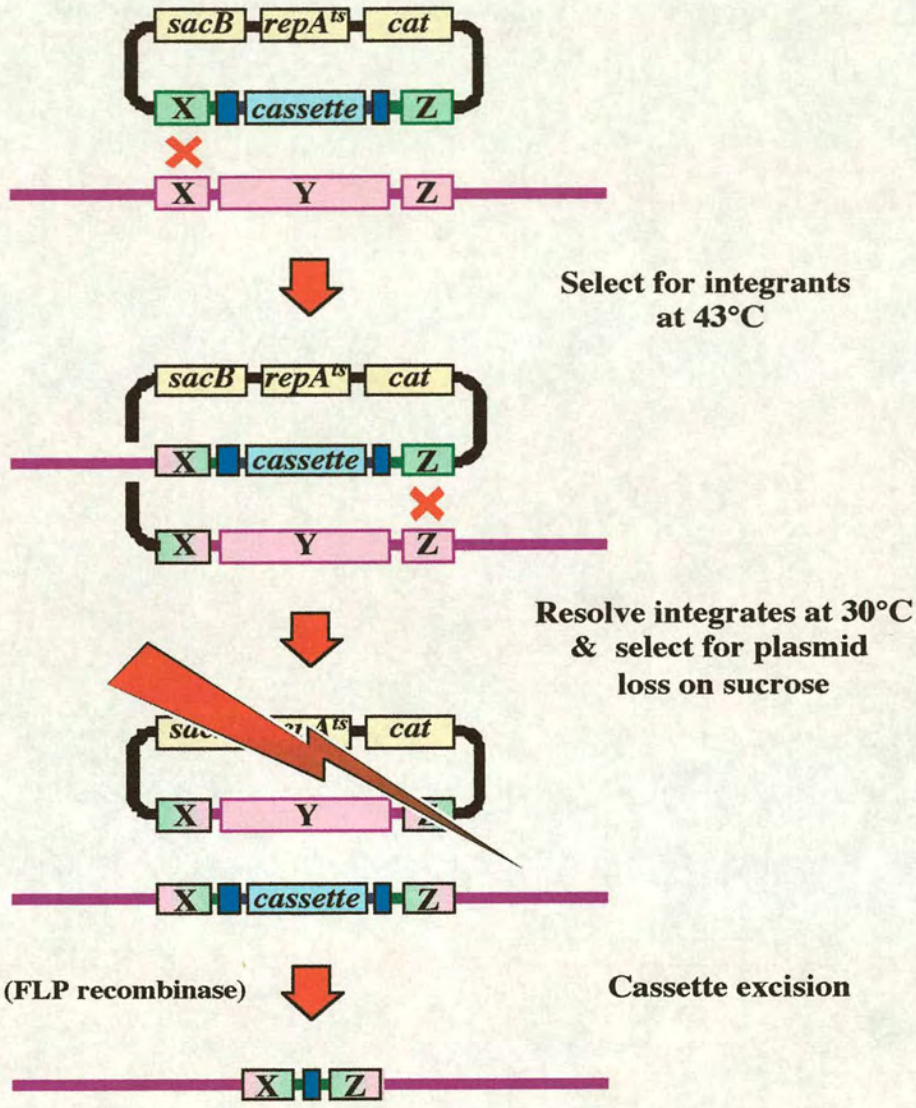


Completed deletion constructs are transformed into a Δ*lac* strain of *E. coli* MG1655 and the replacement procedure begins by streaking transformants on LB chloramphenicol kanamycin plates at 42 deg C. The antibiotics maintain selection of the plasmid and the cloned cassette while the elevated temperature prevents replication of the plasmid pTOF24. The higher temperature therefore selects for plasmid integrants resulting from

a homologous recombination between one of the arms of homology and the corresponding wild-type locus.

After purifying selected integrants, the selective pressure of chloramphenicol and kanamycin resistance is released, allowing the co-integrate to resolve by means of a second homologous recombination. Releasing the plasmid and the second homologous recombination event leaves behind either the wildtype or the mutant copy of the target gene. The deletion procedure selects for the mutant copy of the gene by selecting the engineered kanamycin resistance and for loss of the plasmid borne *sacB* gene by adding sucrose to the growth medium (which proves lethal for strains which have intact *sacB* genes on the plasmid). Loss of chloramphenicol resistance is then screened for.

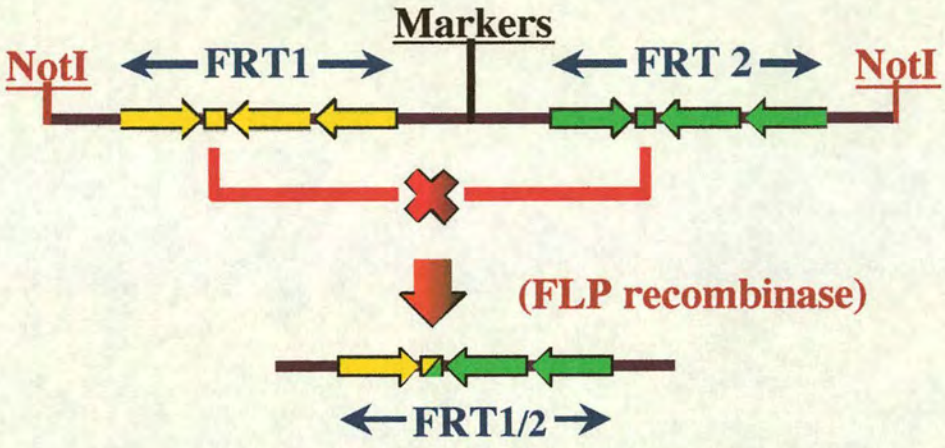
Figure 1.5. Gene replacement with the modified pKO3 method.



Once the target gene is replaced by the mutant allele, native promoter activity and gene expression can be analysed in β galactosidase assays that measure expression of *lacZ* under the control of a native promoter. The FRT (FLP recombinase target) sites flanking the reporter cassette can be used to flip the cassette out to produce a final in-frame deletion (FRT scar -91 base pairs- and truncated ORF) and minimise downstream polar

effects, which might arise from the integrated reporter cassette. This procedure is demonstrated in the cartoon below (figure 1.6.)

Figure 1.6. FLP recombinase and cassette excision.



1.6. Summary

The approach to functionally analysing *E. coli* specific genes adopted in this study was a two step process. The first step was to identify genes which are specific to the *E. coli* group of organisms. This was done using a web based resource at the Microbial Genome Database in Nara, Japan. Precomputed Blast matches of all genes in all gamma proteobacterial genomes on the database were compared and a list of genes specific to the four *E. coli* genomes on the database was downloaded. This web based resource was used to answer questions about *E. coli* specific genes such as, how many genes are specific to the four *E. coli* genomes compared to the 29 other gamma proteobacterial genomes? Has the number of *E. coli* specific genes remained constant since the publication of the first *E. coli* genome? How does altering search parameters (for example Blast cutoff values) affect the number of genes that appear to be specific to *E.*

coli? Does the number of *E. coli* specific genes change by changing which *E. coli* genomes are selected as representative of the entire species?

The second step was to delete a selection of the *E. coli* specific genes from the chromosome of the model *E. coli* strain K-12 MG1655. Genes on the chromosome were replaced with a *lacZ-Km*^{R} cassette flanked by FRT sites as described by Merlin et al, 2002. The mutants were used to study gene expression during growth in LB medium at 37 deg. C. by measuring the levels of ß-galactosidase produced under the control the promoter of the deleted gene. The growth curves obtained were used to assess any differences in growth rate that may have been caused by gene deletion. The mutants were then tested using a variety of different growth conditions (temperature, pH, osmolarity, metal ion toxicity, anaerobiosis and dye toxicity) on LB or minimal agar to detect any phenotypic effects of gene deletion. The observations from growth and expression studies, phenotype analysis and their implications are discussed.

2. Materials and methods.

2.1. Strains.

Table 2.1. List of strains used in this study, their genotypes and sources.

Strain name	*E. coli* genotype	Source
MG1655	F⁻, λ⁻, rph-1.	M. Berlyn.,
EDCM367	MG1655 *ΔlacZ.*	C. Merlin.
DH5α	F⁻, λ⁻, *endA1, hsdR17, hsdM⁺, supE4, thi1, recA1, gyrA96, relA1,* Δ(*argF, lacZYA*)U169, ϕ80D, Δ(*lacZ*)M15.	C. Merlin.
EDCM70	DH5α/pTOF24: DE(*argF lac*)169, phi80dlacZ58(M15), *glnV44*(AS), λ⁻, *rfbD1*, *gyrA96*(*NalR*), *recA1, endA1, spoT1, thi-1, hsdR17.*	C. Merlin.
EDCM81	TOP10/pTOF30 (FLK2 cassette).	C. Merlin.
EDCM145	DH5α/ pTOF73 (FLKP2 cassette).	C. Merlin.
BT340	DH5α/pCP20	B. Wanner.
EDZT1	EDCM367, *yahLM::(lacZ, kan)*	This study
EDZT2	EDCM367, *yahO:: (lacZ, kan)*	This study.
EDZT3	EDCM367, *ybhC:: (lacZ, kan)*	This study
EDZT4	EDCM367, *ybiJ:: (lacZ, kan)*	This study.
EDZT5	EDCM367, *ybiM:: (lacZ, kan)*	This study.
EDZT6	EDCM367, *yccV:: (lacZ, kan)*	This study
EDZT7	EDCM367, *yceP:: (lacZ, kan)*	This study.
EDZT8	EDCM367, *ycfR:: (lacZ, kan)*	This study.
EDZT9	EDCM367, *ycjT:: (lacZ, kan)*	This study
EDZT10	EDCM367, *yncE:: (lacZ, kan)*	This study.
EDZT11	EDCM367, *ydeK:: (lacZ, kan)*	This study.
EDZT12	EDCM367, *ydgH:: (lacZ, kan)*	This study.

EDZT13	EDCM367, *ydhR-Z:: (kan)*	This study.
EDZT14	EDCM367, *ydhV:: (lacZ, kan)*	This study.
EDZT15	EDCM367, *yedJ:: (lacZ, kan)*	This study.
EDZT16	EDCM367, *yegI:: (lacZ, kan)*	This study.
EDZT17	EDCM367, *yegR:: (lacZ, kan)*	This study.
EDZT18	EDCM367, *yeiN:: (lacZ, kan)*	This study.
EDZT19	EDCM367, *yfbL:: (lacZ, kan)*	This study.
EDZT20	EDCM367, *yfjP:: (lacZ, kan)*	This study.
EDZT21	EDCM367, *ypjC:: (lacZ, kan)*	This study.
EDZT22	EDCM367, *ygaQ:: (lacZ, kan)*	This study.
EDZT23	EDCM367, *yqhG:: (lacZ, kan)*	This study.
EDZT24	EDCM367, *ygiN:: (lacZ, kan)*	This study.
EDZT25	EDCM367, *ygjMN:: (lacZ, kan)*	This study.
EDZT26	EDCM367, *yraQ:: (lacZ, kan)*	This study.
EDZT27	EDCM367, *yhcN:: (lacZ, kan)*	This study.
EDZT28	EDCM367, *yhiM:: (lacZ, kan)*	This study.
EDZT29	EDCM367, *hdeB:: (lacZ, kan)*	This study.
EDZT30	EDCM367, *hdeA:: (lacZ, kan)*	This study.
EDZT31	EDCM367, *yigE:: (lacZ, kan)*	This study.
EDZT32	EDCM367, *yihR:: (lacZ, kan)*	This study.
EDZT33	EDCM367, *htrC:: (lacZ, kan)*	This study.
EDZT34	EDCM367, *yjdA:: (lacZ, kan)*	This study.
EDZT35	EDCM367, *ΔyjdI-K:: (lacZ, kan)*	This study.
EDZT36	EDCM367, *ΔyjfY:: (lacZ, kan)*	This study.
EDZT37	EDCM367, *ΔyjiW:: (lacZ, kan)*	This study.
EDZT38	EDCM367, *ΔhtrC:: (lacZ, kan)*	This study.
CA8000	λ^-, e14$^-$, *relA*1, *spoT*1, *thi*$^-$1.	M. Berlyn.
EDZT39	EDCM367, *ΔycfR, ΔyahO.*	This study.
EDZT40	EDCM367, *ΔyhcN::*FLK2, *ΔycfR, ΔyahO*	This study.

EDZT41	EDCM367, *ΔyhcN, ΔyahO, ΔycfR.*	This study.
EDZT42	EDCM367, *ΔydgH::*FLKP2, *ΔyahO, ΔycfR.*	This study.
EDZT43	EDCM367, *ΔydgH, ΔyhcN, ΔyahO, ΔycfR.*	This study.
EDZT44	EDCM367, *ΔybiJ::*FLK2, *ΔydgH, ΔyhcN, ΔyahO, ΔycfR.*	This study.
EDZT45	EDCM367, *ΔyjfY, ΔybiJ, ΔydgH, ΔyhcN, ΔyahO, ΔycfR.*	This study.
EDZT46	EDCM367, *ΔybiM, ΔyjfY, ΔybiJ, ΔydgH, ΔyhcN, ΔyahO, ΔycfR.*	This study
EDZT47	EDCM367, *ΔykgI::*FLKP2, *ΔybiM, ΔyjfY, ΔybiJ, ΔydgH, ΔyhcN, ΔyahO, ΔycfR.*	This study.
206	Ts, *aph*$^+$	S. Raina.
280	Ts, *aph*$^+$.	S. Raina.
CAG12185	λ$^-$, *rph-1, argE86::Tn10.*	M. Berlyn.
EDZT48	CA8000 (Beckwith-CGSC), P1 trnasduced *ΔhtrC, lacZ+, aph+*	This study.
EDZT49	CA8000 (S. Raina), P1 trnasduced *ΔhtrC, lacZ+, aph+*	This study.
TA3515	λ$^-$, *DE(his-gnd)861, ackA202, hisJo-701.*	M. Berlyn.
LCB900	*thr-1, leuB6(Am), fhuA21, lacY1, glnV44(AS),* λ$^-$, *ana-1, rpsL175(strR), thi-1.*	M. Berlyn.
LCB90	*thr-1, leuB6(Am), fhuA21, glnV44(AS), e14-, ana-1, ackB50, nirG.*	M. Berlyn.
78	LCB90, *acetate*$^+$, *aph*$^+$	This study.
80	LCB90, *acetate*$^+$, *aph*$^+$	This study.
90	LCB90, *acetate*$^+$, *aph*$^+$	This study.
EDZT50	EDCM367, (pBAD18-*Cm-yigE*)	This study.
EDZT51	EDCM367, (pBAD18-*Cm*)	This study
EDZT52	EDCM367, *ΔyigE, lacZ+, aph+,* (pBAD18-*Cm-yigE*).	This study
EDZT53	EDCM367, *ΔyigE, lacZ+, aph+,* (pBAD18-*Cm*).	This study.
EDZT54	EDCM367, *ΔyigE*	This study.

CAG12151	λ-, *zdj-925::Tn10, rph-1.*	M. Berlyn.
CAG18464	λ-, *zdj-276::Tn10, rph-1.*	M. Berlyn.
CAG18465	λ-, *zdj-225::Tn10, rph-1.*	M. Berlyn.

2.2. List of plasmids used in this study.

Table 2.2. Details of plasmids used in this study, their characteristics and sources.

Name	Description	Source
pCP20	Ap Cm *repA*(Ts); pSC101-based vector expressing the Flp recombinase.	B. Wanner
pTOF24	pK03 [*Hinc*II-*Hinc*II: 1,252 bp, *aph* from pUC4K]; Cm Km T^s Suc^s	C. Merlin*.
pTOF30	pTOF27 [*Hinc*II-*Hinc*II: 1,252 bp, *aph* gene from pUC4K]; Ap Km	C. Merlin*.
pTOF73	pTOF70 [*Hinc*II-*Hinc*II: 1,252 bp, *aph* gene from pUC4K] Ap Km	C. Merlin*.
pZT1	*YigE* cloned into pBAD-18-Cm.	This study.
pBAD-18-CM		Guzman et al.

*Merlin et al, 2002.

2.3.List of primers used in this study.

Primers No, Ni, Ci and Co used for deleting genes were designed as illustrated in figure 1.2. Check primers used to confirm deletions in some cases were designed upstream and downstream of No and Co primers respectively.

Table 2.3. Details of primers used in this study.

Primer name	Sequence
YkgI:	
Pst I NoykgI	aaaaactgcagcgtgatgagttactgggtgaaggc
Not I NiykgI	cgctcttgcggccgcttggaacggcataacgccagataatagtgttgc
Not I CiykgI	ccgttccaagcggccgcaagagcgtcaaccgcaacagcagtattgtat
Sal I CoykgI	aaaaagtcgacgagtatctttcaataccaggcgac
YahLM	
PstI NoyahLM	aaaaactgcagcagtagctgcgccacataatctcg
NotI NiyahLM	cgctcttgcggccgcttggaacggcggagcttttaatgatatcatcgc
NotI CiyahLM	ccgttccaagcggccgcaagagcgcggcaaatcaggttgattacgtag
SalI CoyahLM	aaaaagtcgacgccaatctctttgtggtagtacaa
YahO	
PstI NoyahO	aaaaactgcagctgcgtaattagcgttgcaagacc
NotI NiyahO	cgctcttgcggccgcttggaacggggtaacggctaacgctaacgcacc
NotI CiyahO	ccgttccaagcggccgcaagagcggacaataagatccacggcacggca
SalI CoyahO	aaaaactcgagctgctgcattacgccaggggttac
NcheckyahO	gtagtgtgctcatttgcggtgttg
CcheckyahO	ggaagaagatatgcagcaaggacg
YbhC	
PstI NoybhC	aaaaactgcagcatgtctttaacggcatgggttac
NotI NiybhC	cgctcttgcggccgcttggaacggaatgccagcgccagacgggaaact
NotI CiybhC	ccgttccaagcggccgcaagagcgaaagtggttgcagaggcgaagaag
SalI CoybhC	aaaaagtcgaccatggtctatgcggtgaaaccagg
YbiJ	

PstI NoybiJ	aaaaactgcaggatgaatgtgaagagtgcggtgcc
NotI NiybiJ	Cgctcttgcggccgcttggaacggaagagccatagcagcaacaacagt
NotI CiybiJ	ccgttccaagcggccgcaagagcgagcggtactgcggtaatttataag
SalI CoybiJ	aaaaagtcgacgcctcgccgcatgatgatgataag
YbiM	
PstI NoybiM	aaaaactgcagggcaaagtctacgctttgctcccc
NotI NiybiM	cgctcttgcggccgcttggaacgggacggtggcaatcagtagtgtgag
NotI CiybiM	ccgttccaagcggccgcaagagcgatgttcggtactgcaaccatctac
SalI CoybiM	aaaaagtcgaccgaaggcgttgatgcttcgacaac
NcheckybiM	gggatatcccagttcgctac
CcheckybiM	tcggatcggcagattgttgtg
YccV	
PstI NoyccV	aaaaactgcaggagctgaacaaactggatctgagc
NotI NiyccV	cgctcttgcggccgcttggaacgggaatttgctggcaatcatagtcac
NotI CiyccV	ccgttccaagcggccgcaagagcgctccaggccccgcgtctgcgtaac
SalI CoyccV	aaaaagtcgacggcattttaacgtccactcacacc
NocheckyccV	gttgcagccaggttgtcagcgttg
CocheckyccV	cccgcttttgctctcattcattcg
YceP	
PstI NoyceP	aaaaactgcagtatgcgtttcctgataatgaagg
NotI NiyceP	cgctcttgcggccgcttggaacgggctgatgtcccaccctacgagcgg
NotI CiyceP	ccgttccaagcggccgcaagagcgtccggtgatttccaggtaaacgag
SalI CoyceP	aaaaagtcgaccagccgggccaaggttaatgacgc
NocheckyceP	ccagctggggctattgacgccctg
CocheckyceP	cattgttcggcgctgtgggcgacg
YcfR	
PstI NoycfR	aaaaactgcaggctccggtcgcttcgacgaggtcc
NotI NiycfR	cgctcttgcggccgcttggaacggcgccgcagcgatgagggtttttac
NotI CiycfR	ccgttccaagcggccgcaagagcgccgaataccctccatggaacagca
SalI CoycfR	aaaaagtcgacgtaagttctggctgtattcgtctg
NoycfRcheck	gagaacggcacgaaataacccctc
CoycfRcheck	tgcctgcggtagtgccggctggac

YcjT	
PstI NoycjT	aaaaactgcaggctggggtgtctttaccaataaag
NotI NiycjT	cgctcttgcggccgcttggaacgggtgataacgttactggcctggtcat
NotI CiycjT	ccgttccaagcggccgcaagagcggctaccaaacatcaggaggatgaa
SalI CoycjT	aaaaagtcgaccctgcgcatcttcaatgccgatac
YncE	
PstI NoyncE	aaaaactgcaggcgaataccgcgaataccgtaagt
NotI NiyncE	cgctcttgcggccgcttggaacggcagatgacgtaaatgcatgacgac
NotI CiyncE	ccgttccaagcggccgcaagagcggatgtgattcgtattgcgctgtaa
SalI CoyncE	aaaaagtcgaccgaaaatgagtcgtcagcatgtgc
YdeK	
NsiINoydeK	aaaaaatgcatgactgacgacgtcgttatggaaag
NotINiydeK	cgctcttgcggccgcttggaacgggcaattccatatcacgcgatagat
NotICiydeK	ccgttccaagcggccgcaagagcgtgtcgctaaaaccactatcggcgg
SalICoydeK	aaaaagtcgacgaatctgtatactgcggctcaccc
YdgH	
PstI NoydgH	aaaaactgcaggcctgtgtcaggagccacacaagc
NotI NiydgH	cgctcttgcggccgcttggaacggtgccgacgccaggagggtgttctt
NotI CiydgH	ccgttccaagcggccgcaagagcgctgaccgtcagcgcagatctgtat
SalI CoydgH	aaaaagtcgacgcgcacagccagtatccccatgcg
NcheckydgH	gccaattcgcatgatattccc
CcheckydgH	accagataggagacgttggcg

YdhRSTUXWVZY combined deletion	
Pst I leftoutside	aaaaactgcagtgcacgttcccgtcgctcatcaag
Not I leftinside	cgctcttgcggccgcttggaacggaagtggtttaagctgctcagccat
Not I rightinside	ccgttccaagcggccgcaagagcggtacagttcatcctcttttgtccg
Sal I rightoutside	aaaaagtcgaccagacgcataacgttcatgccagc
ydhR-Z leftoutside check	caactcgccttcagtaagttg
ydhR-Z rightoutside check	ctgcatagtcaccatgagagaag
YdhV	
PstI NoydhV	aaaaactgcaggggttgccaccgctgtgaaatctc
NotI NiydhV	cgctcttgcggccgcttggaacggattacctgtccaaccgttagccat
NotI CiydhV	
SalI CoydhV	ccgttccaagcggccgcaagagcgctggcagcacacaatctactgcct
NcheckydhV	aaaaagtcgacgcgactttgtaaggacgaggatac
CcheckydhV	gctggtctaacaattgcccct
	cttcctggaagtaatatggcg
YedJ	
PstI NoyedJ	aaaaactgcagctaataacggaagcatcatgacac
NotI NiyedJ	cgctcttgcggccgcttggaacggtgcctgccagtgttgtaagtccat
NotI CiyedJ	ccgttccaagcggccgcaagagcgaaggtgatagatgcgttttcatcc
SalI CoyedJ	aaaaagtcgaccgaaaatggctgacaagggaaacc
YegI	
PstI NoyegI	aaaaactgcagccactcctggccatattctacgtt
NotI NiyegI	cgctcttgcggccgcttggaacggggtcaattcacctgtcgatgtaaa
NotI CiyegI	ccgttccaagcggccgcaagagcggatttaagccgctgctgagccatt
SalI CoyegI	aaaaagtcgacctgatgccactgctgggtatgggt
YegR	

PstI NoyegR	aaaaactgcagacgccgtagaaacttcattaatgg
NotI NiyegR	cgctcttgcggccgcttggaacggataccaacgacactctctagtgtt
NotI CiyegR	ccgttccaagcggccgcaagagcgctgacattcaaatgtggagagtga
SalI CoyegR	aaaaagtcgacctttgtcgtgtcagtgatatagc
YeiN	
PstI NoyeiN	aaaaactgcaggcaacgctattacagctgaatatc
NotI NiyeiN	cgctcttgcggccgcttggaacggagacattctgcgttctccactaac
NotI CiyeiN	ccgttccaagcggccgcaagagcgaaagaatatcagcgtctcgcgggt
SalI CoyeiN	aaaaagtcgacctcattttgcccgaggaaaatagt
YfbL	
PstI NoyfbL	aaaaactgcagctaaacgtcgccaacggattttgc
NotI NiyfbL	cgctcttgcggccgcttggaacgggtaatgcattgacccagggttggt
NotI CiyfbL	ccgttccaagcggccgcaagagcgatggctcaggtagtggatggtgtt
SalI CoyfbL	aaaaagtcgaccaagcatgatcgaggcaaagagtt
YfjP	
NsiI NoyfjP	aaaaaatgcatgtccagagtggaaccgcttcgtta
NotI NiyfjP	Cgctcttgcggccgcttggaacggtttcatcaccactcctctgaaaag
NotI CiyfjP	Ccgttccaagcggccgcaagagcgcgcgctgtctgggatttcttcttc
SalI CoyfjP	aaaaagtcgaccctcaataacctgactcaccacat
YpjC	
PstI NoypjC	aaaaactgcaggggcaaattgcttacatgtggaat
NotI NiypjC	cgctcttgcggccgcttggaacggaactgcgctagcgttaaatccgtt
NotI CiypjC	ccgttccaagcggccgcaagagcggccgatcagtttttgaatgcgct
SalI CoypjC	aaaaagtcgactctgtccagaaaagaagcacccct
YgaQ	
PstI NoygaQ	aaaaactgcagtgagtcgcctgctctaaccactga
NotI NiygaQ	cgctcttgcggccgcttggaacgggataggtaaatttctgggtcctgg
NotI CiygaQ	ccgttccaagcggccgcaagagcgccagacaggataggagaaagaaaa
SalI CoygaQ	caaaagtcgactgcaaatctcccattggattatgg
YqhG	

PstI NoyqhG	aaaaactgcagcgcctgtgataaaccagatcgaac
NotI NiyqhG	cgctcttgcggccgcttggaacggtgccagggctgctaaaaacagaag
NotI CiyqhG	ccgttccaagcggccgcaagagcgaaaactgccgccagattaaagcaa
SalI CoyqhG	aaaaagtcgacgtgcctgcgccaaccattgaagag
YgiN	
NsiI NoygiN	aaaaaatgcatactccaatggtcaactgaacgaca
NotI NiygiN	cgctcttgcggccgcttggaacgggatttctgcgattacggtaagcat
NotI CiygiN	ccgttccaagcggccgcaagagcgaatatccgtattctgcagccaggg
SalI CoygiN	aaaaagtcgaccggtagcgatcggtatcgatcctg
YgjMN	
PstI NoygjMN	aaaaactgcaggttggatcgctgcgagtttatgat
NotI NiygjMN	cgctcttgcggccgcttggaacgggtgcattcttcatcacgtccgttg
NotI CiygjMN	ccgttccaagcggccgcaagagcggccttgtttattgattaacgcgac
SalI CoygjMN	aaaaagtcgacggattggttttgatacggcgatgt
YraQ	
PstI NoyraQ	aaaaactgcaggtttaatccccatgatccgcaact
NotI NiyraQ	cgctcttgcggccgcttggaacggagactgaccagtcatagcattccc
NotI CiyraQ	ccgttccaagcggccgcaagagcgctggcgctgttgttctgatttgag
SalI CoyraQ	aaaaagtcgaccagtgtcggcacccaggtggacga
YhcN	
PstI NoyhcN	aaaaactgcagcgctggcgatgacaccatctttac
NotI NiyhcN	cgctcttgcggccgcttggaacggtagctgggcctttgtttcgtgacc
NotI CiyhcN	ccgttccaagcggccgcaagagcgggtgacacctggcacgctacggct
SalI CoyhcN	aaaaagtcgacgcagaactcgccgcgaaacgtgac
YhiM	
PstI NoyhiM	aaaaactgcagcagtagtattgtcggcgaatctat
NotI NiyhiM	cgctcttgcggccgcttggaacggcatagcataactatagcggggttt
NotI CiyhiM	ccgttccaagcggccgcaagagcgtcaatattagaagcgggttctgct
SalI CoyhiM	aaaaagtcgacctgattagcacattgaccgactgg
HdeB	

PstI NohdeB	aaaaactgcagccaggttataacctcagtgtcg
NotI NihdeB	cgctcttgcggccgcttggaacgggtgacaaagccgctacagcgccca
NotI CihdeB	ccgttccaagcggccgcaagagcgaatcaagcatctaatgacttgccg
SalI CohdeB	aaaaagtcgacgctgcttaaaccagtcgagcattg
HdeA	
PstI NohdeA	aaaaactgcagaccggccagaaattatgactgcgg
NotI NihdeA	cgctcttgcggccgcttggaacggaccaccaagaataacgcctaatac
NotI CihdeA	ccgttccaagcggccgcaagagcgaaagttaaaggcgaatgggacaaa
SalI CohdeA	aaaaagtcgaccttgcacctcattaattcggcaag
NcheckhdeAB	caataaataggcgactgcgacg
CcheckhdeAB	tcctgcaacgaaactaaatcag
YigE	
PstI NoyigE	aaaaactgcagcttcaaagaaacgtgccgatgctt
NotI NiyigE	cgctcttgcggccgcttggaacggaccaatgagtagctgatgcgccat
NotI CiyigE	ccgttccaagcggccgcaagagcgggccaatgttgatggaaaacggtg
SalI CoyigE	aaaaagtcgacgtacgccgatggtcgagaacgaca
SalI NtermyigE	
HindIII	aaaaagtcgacctataatggcgcatcagctactc
CtermyigE	
	aaaaaaagcttcaccgttttccatcaacattg
YihR	
PstI NoyihR	aaaaactgcagcaaccaccagcatggctgctagaa
NotI NiyihR	cgctcttgcggccgcttggaacggcttaattaacgacatacagcctcc
NotI CiyihR	ccgttccaagcggccgcaagagcgtcaggaaaaccgcaccgcctgttt
SalI CoyihR	aaaaagtcgacctgatcgtcggcagagatattcag
HtrC	
NsiI NohtrC	aaaaaatgcatcggcatcgttccaggagaccactc
NotI NihtrC	cgctcttgcggccgcttggaacggaccatccggatgtccaaaaggtcg
NotI CihtrC	ccgttccaagcggccgcaagagcgttgagaaataagcagcttcctcag
SalI CohtrC	aaaaagtcgacggttgctgccgcgttaaccgctca
Ncheck htrC	gcgtctctggcaaccgagtccttc
Ccheck htrC	gagttggaacagttctcaccgcac

Ncheck2htrC	gctcatcaggctgtctacgttcag
Ccheck2htrC	gggataagcgctagttacatgc
htrCseqfor	ctaattgagtcaaactcggcaag
htrCseqrev	actcaactatgatagagacg
YjdA	
PstI NoyjdA	aaaaactgcaggcatctttaacgatcagctcgtcg
NotI NiyjdA	cgctcttgcggccgcttggaacggctcatacagggtctgtgtgtacat
NotI CiyjdA	ccgttccaagcggccgcaagagcgcttttcacggcagaacgatattga
SalI CoyjdA	aaaaagtcgaccagcgcaactcctccagaacgata
YjdI, yjdJ, yjdK	
NsiI NoyjdI-K	aaaaaatgcatgccattcagcgctttaaggatgtc
NotI NiyjdI-K	cgctcttgcggccgcttggaacggcccgtccagtagcgcctgatccat
NotI CiyjdI-K	ccgttccaagcggccgcaagagcgttccgtcagttttacaagtagcca
SalI CoyjdI-K	aaaaagtcgacaacgacgcagaagatcaggctgaa
YjfY	
PstI NoyjfY	aaaaactgcaggcggaaggtagtttccagctcatc
NotI NiyjfY	cgctcttgcggccgcttggaacggaagggctaaaacacgactgaacat
NotI CiyjfY	ccgttccaagcggccgcaagagcgcacgccagcgcggatatttatcgc
SalI CoyjfY	aaaaagtcgacgtcgcttggtggcattgtccatac
YjiW	
PstI NoyjiW	aaaaactgcagcctcagcacgaaatgcaatgatga
NotI NiyjiW	cgctcttgcggccgcttggaacggagcgatgcgggtgtgttggcgcac
NotI CiyjiW	ccgttccaagcggccgcaagagcggccggtaaacagaaagtcgcgtaa
SalI CoyjiW	aaaaagtcgactttgggaacaggcgtaataggact
AckB	
IS903KmRight	agccgtttctgtaatgaag
IS903KmR left	gagccatattcaacggga
Leftinsiderep	gcagagcattacgctgacttg
Rightinsiderep	cccttgtattactgtttatgtaagcagacag

2.4. Genome comparisons: Homologous gene clustering.

(http://mbgd.genome.ad.jp/)

Gene clustering was carried out using precomputed homology clusters stored at the Microbial Genome Database for Comparative Analysis (MBGD) at the Human Genome Centre, Institute of Medical Science, The University of Tokyo. Precomputed BLASTn similarities with P values lower than 10^{-2} between all predicted proteins in the genomes are stored in a database. The program that creates the homology clusters has a web interface where variables like the number of organisms (query), homology parameters (maximum BLASTn value, percent identity etc.) can be defined. Once all variables have been set, the program clusters homologous genes from different genomes and presents the clusters as a histogram or a 'gene cluster map'. In this study, default parameters set on the database were used to search for ORFans. BLASTn cut-off values were changed when required and the results presented in chapter 3. The genomes selected for clustering in this study and analysis of ORFans are detailed in Chapter 3. Chapter 3 also contains a more detailed description on the use of the MBGD database.

2.5. Antibiotic solutions.

The routine concentrations for the antibiotics used in this work are shown in table 2.4 below. All antibiotics were used in both complex and minimal media.

Table 2.4. ***Details of routinely used antibiotics:***

Name	Abbreviation	Solvent	Concentration of stock solution (mg/ml)	Final concentration in media (µg/ml)
Ampicillin	Amp	H_2O	100	100
Chloramphenicol	Chl	Ethanol	20	20

Kanamycin sulphate	Kan	H_2O	25	50
Tetracycline hydrochloride	Tet	50% ethanol	10	10

LB medium was supplemented with sucrose to a final sucrose concentration of 5% (wt/vol) for deletions.

2.6. DNA purification.

All plasmid DNA was prepared from overnight cultures with Promega™ Wizard Plasmid Prep (mini) kits according to manufacturers instructions. For pKO3 derivatives (low copy number plasmid) 50 ml of overnight culture was used for plasmid preparation (this is equivalent to 10 ml of overnight culture per column). Chromosomal DNA was prepared using the Biorad™ Aquapure Genomic DNA isolation kit. All PCR products and restriction digest reactions were purified using the Qiagen™ PCR purification kit.

Chromosomal DNA for southern blotting was prepared using the protocol detailed as Miniprep of Bacterial Genomic DNA, unit 2.4.1. in Current Protocols in Molecular Biology, Volume 1 (Wiley & Sons, 1999).

2.7. Determination of DNA concentrations.

DNA concentrations were determined by measuring the absorption of diluted solutions at 260nm. For double stranded DNA, an OD_{260} value of 1.0 represents a DNA concentration of 50 µg/ml. DNA purity can be determined by measuring absorption at 260 nm and 280 nm. Protein free double stranded DNA should give a 260/280 ratio close to 1.8.

2.8. Digestion of DNA with restriction endonucleases.

Endonuclease cutting of DNA was typically performed in volumes of between 20 and 100 µl. These contained the requisite amount of DNA (between 1-10 µg) and the restriction buffers at 1x concentration as recommended by the manufacturer (Boehringer Mannheim or New England Biolabs). Enzymes were added at concentrations recommended by manufacturers and volumes kept below 10% of total digest volumes to reduce star activity (defined as nonspecific endonuclease activity by manufacturer). Digests were incubated as recommended from 2 hours to overnight using the recommended temperature (usually 37 °C.). All digests were cleaned using Qiagen PCR purification kits according to procedures detailed in kits. Fragments of digests were also recovered from 1% TAE agarose gels using Qiagen Gel Purification kits.

2.9. Ligation of DNA.

Ligations of DNA were usually performed in a final volume of 10 µl. These usually contained between 0.5-1 µg total DNA with insert DNA in a 2-20 fold molar excess over the vector DNA, 1x Boehringer Mannheim ligation buffer and T4 DNA ligase. Ligase (0.2 units) was used for the ligation of cohesive DNA termini, and 1 unit of the enzyme for the ligation of blunt ended molecules. The reactions were incubated for at least 12 hours at 16 °C. Between 5 and 10 µl of the reaction mixture was then used to transform competent cells of an appropriate strain of *E. coli*.

2.10. Agarose gel electrophoresis.

Agarose gel electrophoretic analysis of DNA was always performed using TAE buffer. The gels were made up by melting the appropriate amount of agarose (usually between 0.7 and 1%) in 1x TAE buffer using either a microwave oven or a Bunsen burner. Gels were cast in Biorad or Pharmacia gel trays and, once set, the DNA samples containing 1x tracking dye (6x tracking dye is 0.25% bromophenol blue, 0.25% xylene

cyanol and 40% w/v sucrose in H_2O) were loaded into the wells at one end of the tray. Gels were run in Biorad or Pharmacia gel electrophoresis tanks with their surfaces immersed in 1x TAE buffer. Electrophoresis was usually performed overnight at constant current of 25 mA for 0.7% agarose gels. After completion of electrophoresis, gels were stained in water containing 2 µg/ml ethidium bromide for about 1 hour with constant shaking and subsequently destained in fresh water for about 30 minutes. The gel was then photographed over a UV transilluminator connected to a digital CCD detector.

2.11. Southern blotting procedures.

5 µg of chromosomal DNA prepared and quantitated as described elsewhere was digested with appropriate enzymes according to manufacturer's instructions for 6 hours at 37 °C. 50 µl of the mix was then run on a 0.8% TAE agarose gel overnight at 15 V constant voltage in Pharmacia gel tanks. After running the DNA in the gel was denatured in a solution containing 1.5 M NaCl and 0.5 N NaOH for 1 hour. The gel was then rinsed with distilled water and neutralised for an hour in a solution of 1 M Tris and 1.5 M NaCl. The gel was then placed on a plastic platform covered with 2 pre-wetted (distilled water) Whatman™ type 1 filter papers cut to the shape of the gel. A pre-wetted sheet of Nitrocellulose membrane, cut to the size and shape of the gel, was then placed on top of the gel and two more sheets of pre-wetted Whatman™ were placed on top. Any air bubbles were smoothed out using a glass rod to prevent blockages of capillary action. Dry tissue papers were then stacked on top of the whole setup and a closed plastic container carrying a litre of water was placed on top of the tissue papers to maintain pressure. The transfer was allowed to continue overnight. The setup was dismantled the next day and the DNA adsorbed to the nitrocellulose paper was fixed using an automatic Biorad transilluminator.

Probe DNA was quantitated and 25 ng target DNA alongwith 0.25 ng 1kb DNA ladder (Fermentas™) were radiolabelled with P^{32} dCTP using the random priming method of

High Prime™ as recommended by Roche™. Unincorporated P^{32} dCTP was removed using G-25 Sephadex columns as per manufacturer's instructions (Amersham Biosciences™). The volume of probe was adjusted to 50 µl with distilled water and denatured by boiling for 10 minutes. The denatured probe was then added to pre-wet membrane at 65 °C in 20 ml hybridisation buffer and hybridisation allowed to continue overnight. Hybridisation buffer includes 10ml 1 M $NaPO_4$, 7 ml 20% SDS and 3 ml distilled water. After overnight incubation the membrane was washed 3 times in fresh hybridisation buffer at 55 °C. wrapped in cling film and exposed to a phosphor screen overnight (Molecular Dynamics™). The phosphor screen was scanned using a Storm 860 Imager (Molecular Dynamics™). Images generated were processed and analysed using Imagequant (Amersham Biosciences).

2.12. Competent cells for heat shock transformations.

Competent *E. coli* DH5α were prepared using the $RbCl_2$ method. A single overnight colony of DH5α was inoculated into 5 ml LB medium and grown overnight. The overnight culture was diluted 1/150 into two 30 ml aliquots of LB in 250 ml flasks. The cells were allowed to grow to an optical density of 0.4 (4-7 x 10^7 cells at 550nm) while shaking at 300-320 rpm. The culture was transferred to cold 33 ml tubes, chilled on ice for 10 minutes and cells were pelleted at 3000g for 10 minutes (4500 rpm @ 4°C). Supernatant was decanted and the pellet was resuspended in 1/3 vol. (10 ml) RF1, (100 mM $RbCl_2$, 50 mM $MnCl_2$, 30 mM potassium acetate, 10 mM $CaCl_2$, 15% (v/v) glycerol, final pH adjusted to 5.8 with 0.2% acetic acid and filter sterilised with a 0.22 micron filter), by gentle vortexing and incubated on ice for 45 minutes. Cells were pelleted at 3000g (4°C) for 15 minutes, supernatant decanted and cells gently resuspended in 1/12.5 RF2, (100 mM MOPS, 10 mM $RbCl_2$, 75 mM $CaCl_2$, 15% glycerol, pH 6.8 with NaOH, filter sterilise as above) incubated on ice for 15 minutes. 200 µl aliquots were transferred to cold 1.5 ml centrifuge tubes and flash frozen with liquid nitrogen.

Competent *E. coli* MG1655 for gene deletions were prepared fresh for each transformation by the $CaCl_2$ method detailed below. Overnight cultures of the strains (grown without antibiotic selection) were inoculated (1:50, v/v) into fresh LB medium (5 ml). The culture was grown to an optical density (OD) between 05-0.6 at 540 nm. The culture was chilled on ice and transferred to cold 2 ml centrifuge tubes. Cells were pelleted at high speed in an MSE Micro Centaur microcentrifuge for 25 seconds. Supernatant was decanted and pellet was resuspended in 1/2 the original volume of ice cold 0.1 M $MgCl_2$. Cells were pelleted again as above, supernatant decanted and the pellet resuspended in 1/20th the original volume of ice cold 0.1 M $CaCl_2$. This cell suspension was incubated on ice for 2 hours prior to heat shock transformations.

2.13. Heat shock transformations.

If cells were frozen they were thawed on ice. 5 µl of plasmid DNA (freshly prepared or from overnight ligations) were added to 50-100 µl of cells and incubated on ice for 30 minutes. For $RbCl_2$ frozen competent cells, heat shock at 42 °C was carried out for 1 minute. For cells prepared with the $CaCl_2$ method heat shock was continued for 2 minutes followed by 2 minutes on ice. Following this, cells were incubated at 30 °C in 900 µl of LB medium for 2 hours before being spread on selective plates.

2.14. In-vitro deletion construction.

All PCR reactions were performed in a Hybaid™ PCR *Sprint* machine. To prevent possible errors during PCR, the 3' → 5' proof reading DNA polymerase, *Pfu* (Promega UK), was used at a concentration of 1 unit per 50 µl PCR reaction. In addition, each PCR reaction also had 1 unit of *Taq* polymerase (Roche) added for increased efficiency. These conditions were maintained for all PCR reactions and not just for deletion construction. The primers were designed so that both arms (N and C terminal arms with respect to the target gene) would be between 450-500 base pairs (bp.) providing a total homology of about 0.9 to 1 kbp to the flanking regions of the ORF to

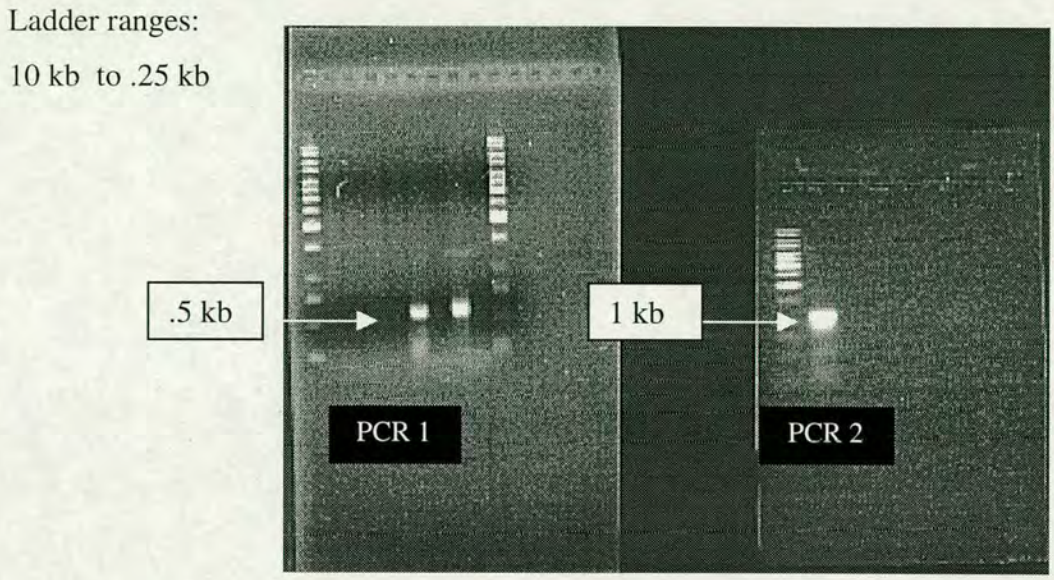
be deleted. All inside primers were designed in-frame with the predicted start and stop codons of the ORF to be deleted (figure 1.2.). This was done to prevent any possible frame shifts when the reporter cassette was later removed from the chromosome.

All primers were used in 10 μM concentrations. The N outside (No -respective to the N terminal of ORF to be deleted-) primers carried a *Pst*I or compatible sites, whereas the C outside (Co - respective to the C terminal of ORF to be deleted-) primers carried a *Sal*I or compatible site, for ligation into the corresponding *Pst*I, *Sal*I sites of pTOF24. Both the N and C inside (Ni and Ci) primers carried a 'bridge' sequence (ccgttccaagcggccgcaagagcg) which made the crossover PCR and incorporation of *Not*I restriction site possible.

The two flanking arms of the construct were amplified seperately in reactions labelled PCR 1. The two arms were then used as a templates in a second reaction labelled PCR 2 were they were fused. PCR 1 was carried out in a 50 μl reaction with a 250-500 μg/μl dilution of MG1655 genomic DNA. The PCR reaction mix was denatured for 1 min. at 94°C followed by 30 cycles of denaturation (1' at 94°C), annealing (30'' at 55°C) and extension (1' at 72°C). The final extension was carried out for 5' at 72°C. PCR products (15-20 μl) were run on 1% agarose (Sigma) TAE gel to confirm product size and abundance.

PCR 2 was carried out in a 100 μl reaction with appropriate volumes of PCR 1 products (1-3 μl) used as template (cleaned with Qiagen PCR purification kit if appropriate). The extension time for the crossover PCR was extended to 2 minutes and the annealing temperature varied between 55°C to 63°C depending on the GC content of the primer in question. The program otherwise was unchanged from the one detailed for PCR 1 (above). Products were run on 1% agarose (Sigma) TAE gel to confirm size and abundance and purified using Qiagen's PCR purification kit.

Figure 2.1. PCR 1 and 2 (crossover) of the htrC *deletion.*



Cloning PCR deletion products: The crossover PCR product was cut at its ends with 10 units of *Pst*I and *Sal*I (or compatible enzymes, supplied by Boehringer Mannheim) for 2 hours (or overnight) in a 50 µl reaction. The vector (pTOF24) was prepared from a 50 ml overnight culture (Promega Wizard) and bulk digested overnight with *Sal*I and *Pst*I in a 200 µl digestion mix. The DNA vector was then de-phosphorylated with 16 units of calf intestinal alkaline phosphatase (Roche Molecular Biochemicals) for 1 hour at 37°C before being cleaned as before. The vector and crossover PCR product were then fused in an overnight 20 µl ligation reaction at 16 °C using 2 units of T4 DNA ligase (Roche Molecular Biochemicals).

Competent *E. coli* DH5α was transformed with 5 µl of the ligation reaction and cells were plated on LB chloramphenicol. All plates were incubated at 30°C, unless mentioned otherwise, to maintain the temperature sensitive plasmid. Chloramphenicol resistant colonies (50) were patched on LB plates containing both chloramphenicol and kanamycin and on plates containing chloramphenicol, as a plasmid digestion control. Colonies exhibiting the desired phenotype (Cp^R, Km^S) were analysed further by

extracting plasmids (Promega wizard mini kit) and digesting them with *SalI* and *PstI*. Insert sizes were confirmed on 1% agarose gels in TAE. Clones with the right size insert were then used to clone the reporter cassette (figure 2)

Figure 2.2. Crossover cloning analysis.

Lanes 1-6 show digested vector (~6 kb) and inserts (~900 bp), lane 7 carries the 1Kb ladder by Fermentas™. Fragment lengths of ladder are as defined by manufacturer.

The reporter cassettes FLK and FLKP2 were excised as *Not*I fragments from pUC18 based plasmids pTOF30 and pTOF72 (courtesy Dr. C. Merlin). The *lacZ* gene was used in studying the expression of the remaining promoters of deleted genes while kanamycin resistance provided a marker to select for clones with integrated cassettes. Flanking the *lacZ* and *aph* genes are FRT sites (FLP recombinase target) which can be used to flip the cassette out using a plasmid that expresses the FLP recombinase protein.

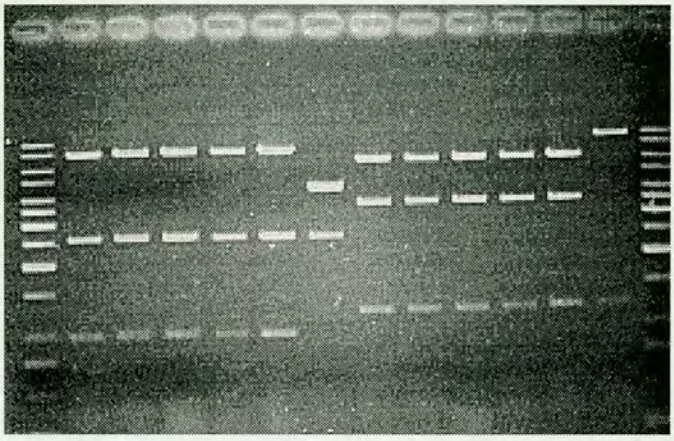
The *NotI* fragment of the reporter cassette was ligated to the *Not*I digested vector (pTOF24 with a cloned crossover PCR product) in a reaction similar to the one described for the crossover PCR ligation. 5 µl of an overnight ligation reaction were used to transform 50 µl of One-Shot TOP10™ (Invitrogen) chemically competent *E. coli* DH5α as per manufacturer's instructions. Colonies showing resistance to kanamycin

and chloramphenicol were analysed for the orientation of the reporter cassette by restriction enzymes. Clones showing the desired orientation (positively oriented *lacZ* and *aph* cassettes) were used for gene replacement.

All orientation analysis experiments were modelled on Mapdraw™ (DNA*™). Enzymes showing maximum resolution were chosen for each analysis. Most often this would be a DNA digestion with enzymes *Hind*III and *Pst*I.

Figure 2.3: Orientation analysis of reporter cassette cloned into the htrC *deletion vector.*

1 2 3 4 5 6 7 8 9 10 11 12 13 14



Lanes 1 & 14 → 1kb DNA ladder (Fermentas™).
Lanes 2-7 → HindIII digest
Lanes 8-13 → PstI digest
Lanes 2-6 & 8-12 → positively oriented reporter cassettes
Lanes 7 & 13 → negatively oriented reporter cassette.

2.15. Gene replacement.

Mutant alleles cloned into the pTOF24 replacement vector were heat shock transformed into competent *E. coli* MG1655 *ΔlacZ* cells as detailed in section 2.13. The resulting chloramphenicol and kanamycin resistant colonies were spread on fresh LB plates containing both antibiotics and incubated at 42°C. At this restrictive temperature (for plasmid replication from the pSC101 temperature sensitive origin of replication) only clones containing pTOF24 plasmids integrated into the chromosome would produce colonies resistant to the two antibiotics. Clones were purified once at 42°C

before selecting a single colony for plasmid excision at 30°C on LB broth overnight followed by serial dilution and plating dilutions on LB and LB plates supplemented with 5% (w/v) sucrose and kanamycin.

Colonies resistant to kanamycin and able to grow in the presence of sucrose (loss of *sacB*) were patched on LB plates containing both chloramphenicol and kanamycin. Clones that were Cp^S and Km^R were analysed for gene replacement with PCR using either No and Co primers or primers homologous to flanking sequences of the ORF replaced (but upstream and downstream of the No and Co primers respectively) and primers specific to the cassette. Most clones that were sensitive to chloramphenicol gave PCR fragments of the expected size. These clones were then used for survival and expression analysis.

2.16. Growth curves and β-galactosidase assays.

The method used here is essentially the same as that described by Miller (1972). All growth curves were carried out in 20 or 50 ml LB broth in 100 ml or 250 ml conical flasks respectively. Most growth experiments included two successive starter cultures before testing expression in a desired condition. This proved necessary since most overnight cultures showed high levels of β-galactosidase activity. Starter cultures were overnight cultures diluted 1:100 in fresh LB medium and grown to an optical density (O.D.) of 0.2 at 600nm and rediluted into fresh LB and grown again to an OD of 0.2 after which the culture was diluted 1:4 into test media. This was used as the starting point for enzyme measurement and 0.8 - 1 ml samples were then taken at set intervals and optical density recorded in plastic cuvettes on a Hitachi U-2000 spectrophotometer at 600nm.

For the β galactosidase assays 0.1 ml of culture was mixed with 0.9 ml Z buffer in small glass test tubes, followed by addition of 50 µl of chloroform (to permeabalize the cell membrane) and vortexing the mix for 15 seconds. Samples were stored at 4°C until

ready to assay. For the assay, 0.1 ml of LB broth mixed with 0.9 ml Z buffer was used as a control. 200 µl of orthonitrophenol-β-d-galactopyranoside (ONPG 4 mg/ml) were added to each tube and mixed by vortexing. All samples were then placed in a water bath at 30°C and the exact time noted. Tubes were checked every 10 minutes and when yellowing was observed 0.5 ml of Na_2CO_3 was added to stop the reaction and the time noted (T). Optical density for all samples was measured at 420 and 550 nm. β galactosidase activity was calculated in Miller units using the formula given below.

$$\beta\text{-galactosidase activity} = \frac{O.D._{420}-(1.75 \times O.D._{550}) \times 1000}{O.D._{600} \times 0.1 \times T}$$

Where

T = time in minutes for color change.

0.1 = sample volume of test culture.

Total enzyme activity was calculated by multiplying β galactosidase activity by the optical density of the culture at 600nm at each sample point.

Z-buffer (per liter): 0.06 M $Na_2HPO_4.7H_2O$ 0.001 M $MgSO_4.7H_2O$

0.04 M $NaH_2PO_4.H_2O$ 0.01 M KCl

0.05 M β-mercaptoethanol

To calculate total expression values, Miller units were multiplied by the appropriate dilution factor and the dilution corrected Miller units were then multiplied by the dilution corrected optical density of the culture (OD 600nm) at sample time. Total expression values are therefore ß-galactosidase levels/ml of culture irrespective of optical density.

2.17. Frozen storage of bacterial strains.

E. coli strains with or without plasmids could be conveniently stored at –20 or –70 deg C. without suffering a dramatic loss of viability. A fresh 5 ml overnight culture

was prepared with antibiotic selection if required. This was centrifuged at 4000 rpm for 10-15 minute, the supernatant discarded and the cells resuspended in 0.1 x the original volume of Frozen Storage Buffer. The cell were left on ice for an hour before storing at –20 or –70 deg C.

Frozen storage buffer:	50% bacterial buffer
	50% glycerol (v/v)

2.18. P1 lysate preparation.

Overnight *E. coli* strains were used to inoculate fresh LB broth with 2.5 mM $CaCl_2$ and were shaken at 37 °C. until the culture reached an OD of 1 at 600nm. 1 ml aliquots were then mixed with 10^{-5}, 10^{-6} and 10^{-7} P1 phage (wildtype MG1655 derived) and incubated at 37 °C, without shaking, in large test tubes for 20 minutes. 1ml molten LC top agar (60 °C) was added and the mixture briefly vortexed before pouring onto LB agar plates. Plates were incubated overnight at 37 °C. without inversion. Top agar was scraped off with 5 ml LB with 2.5mM $CaCl_2$ if visible plaque formation was observed. The mixture was shaken at 30 °C. for 20 minutes with 100 µl of chloroform. The entire mixture was poured into sterile glass universal bottles and centrifuged at 4500 rpm for 15 minutes. The supernatant was stored over a few drops of chloroform in 10 ml universal bottles at 4 °C.

2.19. P1 transduction procedures.

Overnight *E. coli* strains were used to inoculate fresh LB broth and were shaken at 37 °C. until the culture reached an OD of 1 at 600 nm. 10 ml of the culture were centrifuged at 4500 rpm at room temperature in universal bottles and the pellet resuspended in $1/10^{th}$ the volume of LB with 2.5mM $CaCl_2$. 100 µl of 10x concentrated cells were then incubated with 1, 10 and 100 µl of the appropriate phage lysate and incubated at 37 °C for 15 minutes. 800 µl of LB were added to the mix and 250 µl

aliquots were spread on to selective agar media with an appropriate antibiotic or nutritional composition. Plates were incubated at a suitable temperature until colonies appeared.

2.20. Media.

L-broth	Difco bacto tryptone	10 g
	Difco bacto yeast extract	5 g
	NaCl	5 g
	pH to 7.2 with NaOH	
	Distilled water to 1 litre	
L-agar	L-broth + 15 g Difco agar per litre	
M9 minimal medium (4x):	Na_2HPO_4	28 g
	KH_2PO_4	12 g
	NaCl	2 g
	NH_4Cl	4 g
	Distilled water	1 liter
Phage buffer:	Na_2HPO_4	7 g
	KH_2PO_4	3 g
	NaCl	5 g
	$MgSO_4$ (0.1M)	10 ml
	$CaCl_2$ (0.1M)	10 ml
	1% gelatin solution	1 ml
	Distilled water to 1 litre	
Bacterial buffer	KH_2PO_4	3 g
	NaH_2PO_4	7 g

	NaCl	4 g
	$MgSO_{4\cdot7}H_2O$	2 g
	Distilled water to 1 litre	

TE buffer	10 mM Tris-HCl (pH 8.0)	
	1 mM EDTA (pH 8.0)	

TAE buffer	*Working solution:*	
	40 mM Tris-acetate	
	2 mM EDTA	
	50x concentrated stock solution:	
	Tris base	242 g
	Glacial acetic acid	57.1 ml
	0.5 M EDTA (pH 8.0)	100 ml
	Distilled water to 1 litre	

2.21. Composition of media used for phenotypic tests:

Anaerobic growth conditions were maintained by incubating petri plates with LB agar in anaerobic jars supplied by BQ-BBL: GasPack System^{R}. Anaerobiosis was generated using the Oxoid Anaerogen^{TM} system as per manufacturers instructions. Agar plates with different pH's were prepared as detailed below

Stock buffers were mixed to adjust the pH of M9 agar medium as shown below.

- pH 5.6 13.7 ml 0.1 M citric acid
  36.3 ml 0.1 M sodium citrate
  adjust to 100 ml with M9 agar.

- pH 5.8 11.8 ml 0.1 M citric acid

38.2 ml 0.1 M sodium citrate
adjust to 100 ml with M9 agar.

- pH 9 — 50 ml 0.2 M boric acid
  59 ml 0.05 M borax
  adjust to 200 ml with M9 agar.

- PH 9.2 — 50 ml 0.2 M boric acid
  115 ml 0.05 M borax
  adjust to 100 ml with M9 agar.

LB agar plates containing different amounts of metal ions, salt and crystal violet were prepared by diluting 1 M stock solutions of salts $CoCl_2$, NaCl, $ZnCl_2$, $NiCl_2$.6H20, $CuCl_2$ and crystal violet (0.2%).

2.22. λ mini-Tn*10 library preparation and use.*

Preparation of the λ library was carried out as described by Kleckner (1991). The phage used in this study was λNK1316. 10 µl of stock phage were mixed with 150 µl of EDCM367 after the recipient strain had reached an OD_{600} of between 0.7 and 1 in LB broth with 0.1M $MgCl_2$ at 37 °C. The mixture was incubated at 37 °C. for 20 minutes after which the entire mix was poured into 10 ml LB broth with kanamycin and incubated at 37 °C. with shaking for 24 hours. The 24 hour culture was then used to prepare a P1 lysate using a concentration of 10^7 stock phage as described elsewhere.

Chapter 3: Identification and functional analysis of genes unique to *E. coli:*

3.1: Identification of ORFan or species-specific sequences in E. coli K-12 MG1655.

E. coli specific genes in this study were identified using precomputed BLASTn matches stored at the Microbial Genome Database (MBGD; http://mbgd.genome.ad.jp/). This database was chosen since it provides the means to identify ORFan genes in *E. coli* without the need for any special modifications or advanced programming skills on the user's part. According to the database "Genes in the chosen genomes are classified by the hierarchical clustering method known as UPGMA using precomputed similarity relationships identified by all-against-all BLAST searches. The result is displayed as a histogram which we call 'Gene Cluster Map' where clusters obtained are summarized by the phylogenetic patterns (representing presence or absence of each orthologous group in each genome) and function categories."

3.2. Genome comparisons at the MBGD database.

Shown below is the method used to obtain a list of *E. coli* specific genes from the MBGD web site. The MBGD web page has links to sections titled "organism selection on the taxonomy browser" and "create/view orthologous clusters". The first was used to select only the gamma proteobacterial genomes on the database and the second was then used to create a list of orthologous clusters. Figure 3.1 shows the gamma proteobacterial genomes selected for comparisons. The link called 'Create view orthologous genes' then produces the clusters described below.

The orthologous clusters are reported as a graphical chart with clusters shared between genomes represented as green blocks (marked with an asterisk) in figure 3.2. below. The

number of genes in each block, along with their colour-coded functional classifications are summarised in a column to the right of the blocks.

Figure 3.1. The MBGD database; gamma proteobacterial genomes used in this study are underlined.

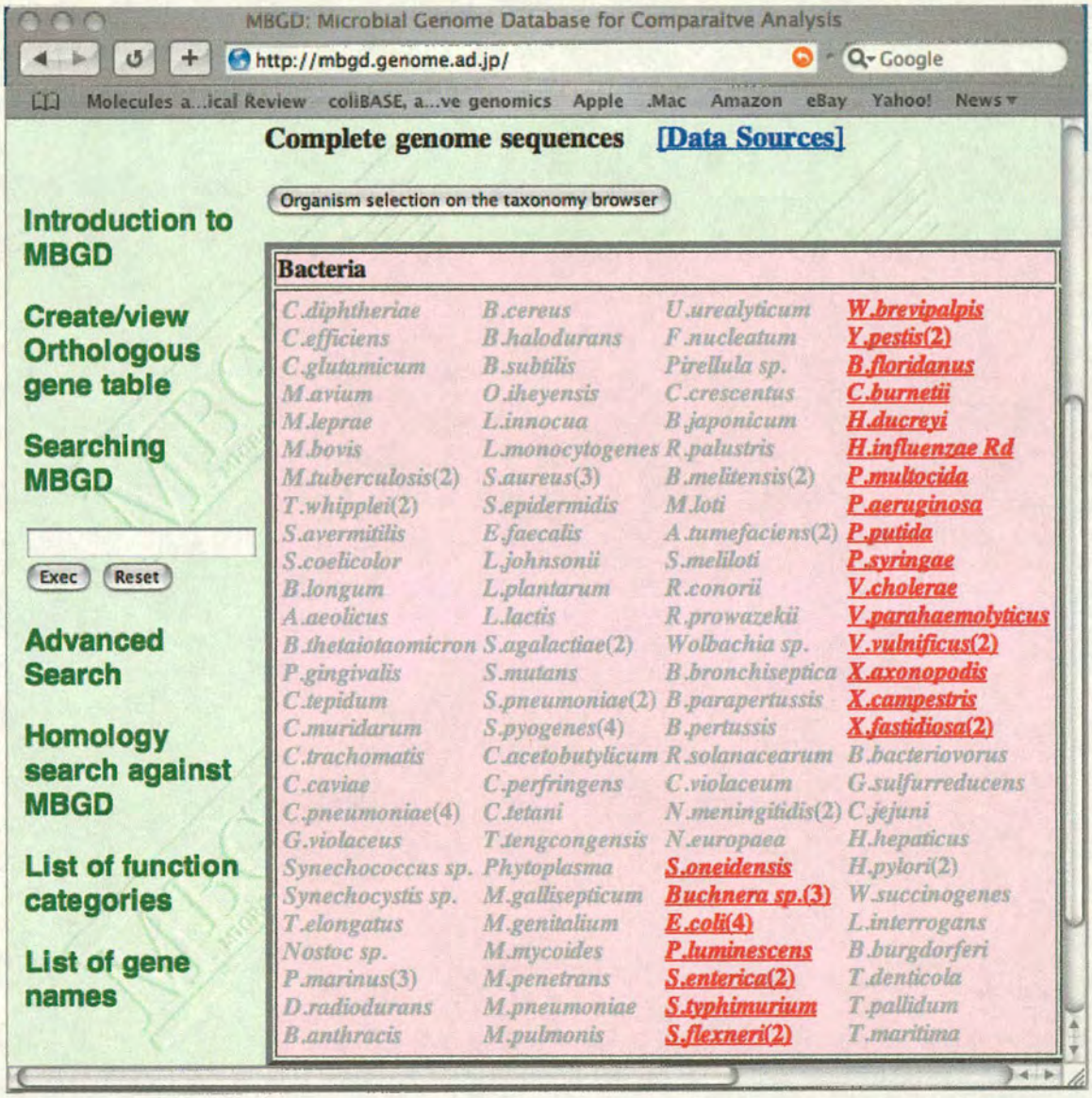
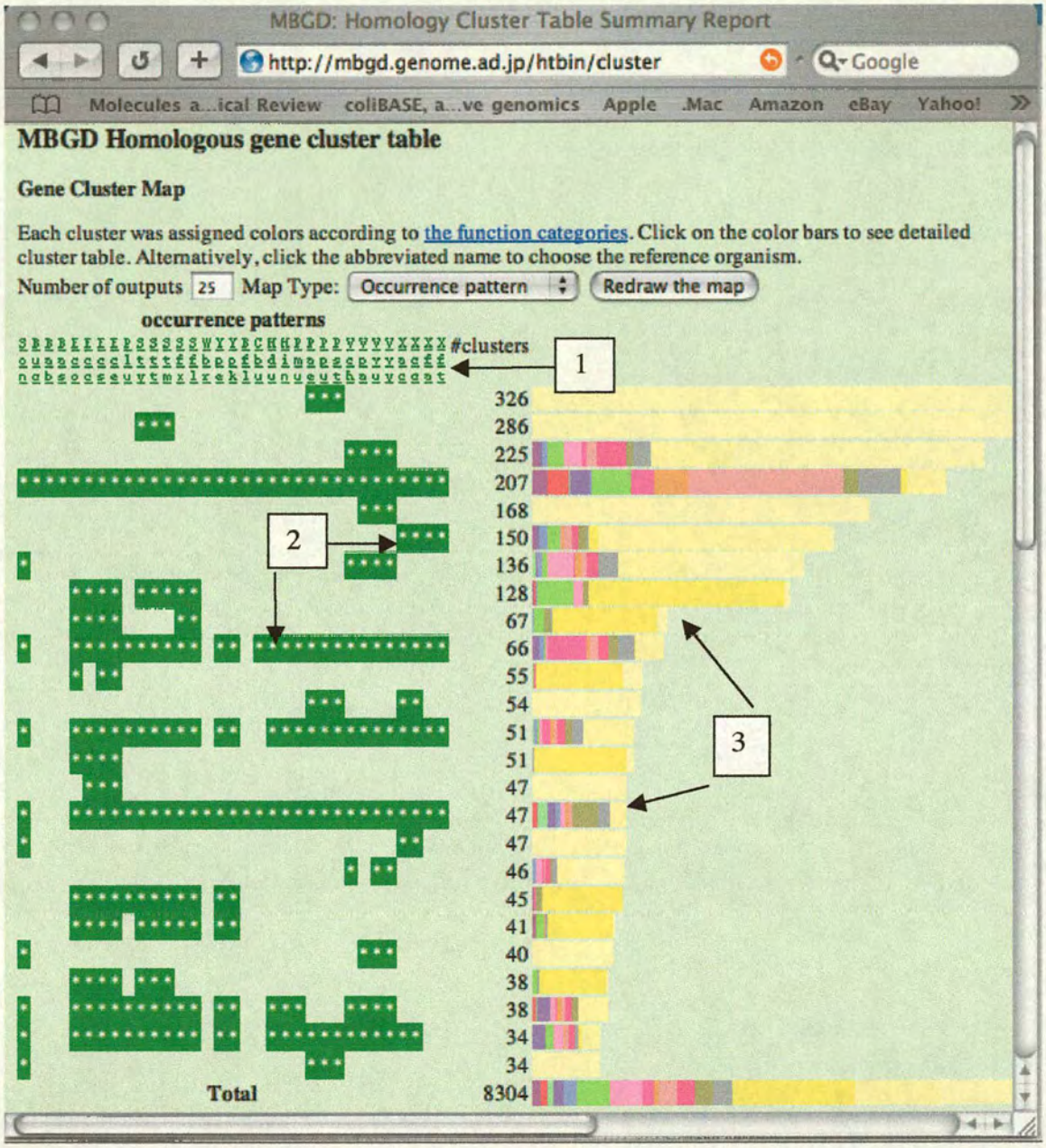


Figure 3.2. The MBGD cluster table.



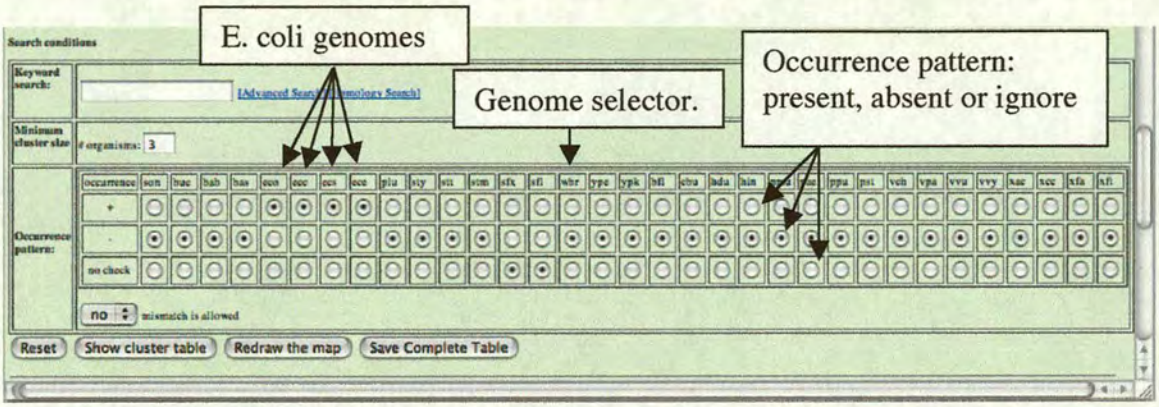
1) *Abbreviated names of organisms used in the cluster analysis.*

1) *Blocks representing shared clusters in different genomes.*

1) *Colour coded distribution of functional classes of gene in each cluster.*

The cluster table was rearranged to list clusters of genes that are present in the four *E. coli* genomes and absent in other gamma proteobacterial genomes using the interactive table shown in figure 3.3. below. The list of clusters was downloaded as a tab delimited text file which can be opened using a program such as Microsoft Excel.

Figure 3.3. Filtering clusters specific to E. coli.



The MBGD database has been regularly updated over the last four years and remains under active development. While allowing the user to download results of pre-computed BLASTn analyses the database also allows users to change homology parameters and to add or delete single or multiple genomes to/from the analysis. It is therefore an easy, configurable tool for identifying ORFan genes within any genome on the database.

3.3. E. coli *specific ORFs in relation to sequenced gamma proteobacteria.*

The four sequenced *E. coli* genomes (K-12 MG1655, 0157:H7, EDL933 and CFT073) were compared to all sequenced gamma proteobacterial genomes. Genes common to the four *E. coli* genomes which have orthologs (default setting $p=10^{-2}$) in sequenced gamma proteobacterial genomes were excluded, leaving a list of genes which were specific to the four *E. coli* genomes. The MBGD database has 33 (as of 19.5.04) fully-sequenced gamma proteobacterial genomes. Below is a list of bacterial genomes used in the cluster analysis.

Table 3.1. Gamma proteobacterial genomes in the order they were sequenced (left-right).

Haemophilus influenzae Rd (Mar-96).	*Escherichia coli* K-12 MG1655 (Sep-97).	*Xylella fastidiosa* 9a5C (Jul-00).
Pseudomonas aeruginosa (Aug-00).	*Vibrio cholerae* (Aug-00).	*Buchnera* sp. APS (Sep-00).
Escherichia coli EDL933 (Jan-01).	*Escherichia coli* 0157:H7 (Feb-01).	*Pasturella multocida* PM70 (Mar-01).
Yersinia pestis CO92 (Oct-01).	*Salmonella enterica* CT18 (Oct-01).	*Salmonella typhimurium* (Oct-01).
Xanthomonas campestris *(May-02).*	*Xanthomonas axonopodis* (May-02).	*Buchnera aphidicola* (Schizaphis graminum) (Jun-02).
Yersinia pestis KIM (Aug-02).	*Shigella flexneri* 2a 301 (Oct-02)	Shewanella oneidensis (Nov-02).
Wigglesworthia brevipalpsis (Nov-02).	*Pseudomonas putida* KT2440 (Dec-02).	*Escherichia coli* CFT073 (Dec-02).
Buchnera aphidicola (Baizongia pistaciae) (Jan-03).	*Xylella fastidiosa* Temecula1 (Feb-03).	*Vibrio parahaemolyticus* (Mar-03).
Salmonella enterica Ty2 (Apr-03).	*Coxiella burnetii* (Apr-03).	*Shigella flexneri* 2a 2457T (May-03).
Candidatus Blochmannia floridanus (Aug-03).	*Photorhabdus luminescens* (In-press).	*Haemophilus ducreyi* (Unpublished).
Pseudomonas syringae pv. tomato (Unpublished).	*Vibrio vulnificus* CMCP6 (Unpublished).	*Vibrio vulnificus* YJ016 (Unpublished).

The two sequenced *Shigella* genomes were included in the cluster analysis but due to the very close phylogenetic relationship with *E. coli,* orthologs common to the four *E. coli's*

and/or *Shigella* were listed as *E. coli* ORFans i.e. ORFs/genes were labelled *E. coli* ORFans if they were conserved in all four *E. coli* genomes and absent in all other gamma proteobacterial genomes not including *Shigella*. Below is a table listing ORFan clusters common to all four *E. coli* genomes and absent in all gamma proteobacterial genomes above, except for *Shigella*, according to the MBGD database. The term 'cluster' in this study indicates either single genes or a family of paralogous genes found to be present in the four *E. coli* genomes and absent in other gamma proteobacterial genomes.

Table 3.2. E. coli *specific clusters (as of 19.5.04)*

Blattner number	**Predicted motifs / or known function**	**Name**	**Length (a.a.)**	**Known function**
B0079		*fruL*	28	no
B0132	Putative transposase – YhgA like.	*yadD*	300	no
B0252 B2633		*yafZ*	278	no
B0289		*yagV*	251	no
B0290	C terminal similar to class C vacuolar protein sorting (Vps) complex	*yagW*	537	no
B0292		*yagY*	222	no
B0293	Serine hydroxymethyltransferase domain	*yagZ*	195	no
B0294	Putative LuxR family regulator protein	*ykgK*	196	no
B0317		*yahC*	165	no
B0319		*yahE*	287	no
B0320	Putative CoA type ligase	*yahF*	515	no
B0321		*yahG*	472	no
B0323	Putative amino acid kinase family	*yahI*	316	no
B0392		*ykiA*	93	no
B0486	Putative prokaryptic lipoprotein, amino acid/amine transporter. Also has AraC like	*ybaT*	430	no

	regulatory domain			
B0542			45	no
B0550	Resolvase; resolves Holliday structures	*rus*	120	yes
B0554 B1556	Putative lysis protein S	*ybcR*	71, 96	no
B0557	Putative prokaryotic lipoprotein	*ybcU*	97	no
B0609	Similar to PapB, Adhesin biosynthesis transcription regulatory protein.		*153*	no
B0771	Putative Aconitase C like enzyme	*ybhJ*	761	no
B0833(2)	Domains DUF2 (conserved eukaryotic domain) and EAL (found in bacterial signal proteins)	*yliE*	782	no
B0843		*ybjH*	94	no
B0991		*sfa*	76	no
B1006	Putative prokaryotic lipoprotein similar to Xanthine Uracil transporters	*ycdG*	464	no
B1198	Phospho carrier like protein	*ycgC*	473	no
B1295		*ymjA*	81	no
B1316	Putative prokaryotic lipoprotein.	*ycjT*	755	no
B1319	Putative prokaryotic lipoprotein.	*ompG*	301	yes
B1415	Aldehyde dehydrogenase, NAD linked	*aldA*	479	yes
B1454(1)	Carries a Glutathione S-transferase, N-terminal domain.		205	no
B1500			65	no
B1502	putative adhesin; similar to FimH protein	*ydeQ*	304	no
B1555		*ydfR*	103	no
B1569	Regulatory for *dicB*	*dicC*	76	yes
B1572		*ydfB*	56	no
B1573		*ydfC*	72	no
B1575	Control of cell division. activates MinC	*dicB*	62	yes
B1576		*ydfD*	63	no
B1615(1)	Membrane-associated protein in *gus* operon	*uidC*	417	yes

B1616	Glucuronide permease	*uidB*	457	yes
B1617(1)	Beta-D-glucuronidase	*uidA*	603	yes
B1668	Phosphomannose isomerase like protein	*ydhS*	534	no
B1669		*ydhT*	270	no
B1671	putative oxidoreductase, Fe-S subunit	*ydhX*	239	no
B1672		*ydhW*	215	no
B1673	Aldehyde ferredoxin oxidoreductase like protein	*ydhV*	700	no
B1674	Putative prokaryotic lipoprotein	*ydhY*	208	no
B1721(2) B4017(3)	Regulator of acetyl-coenzyme A synthetase gene expression	*arp*	471	yes
B1751	Putative prokaryotic lipoprotein	*ydjY*	279	no
B1770	Putative regulator of the DeoR type family	*ydjF*	252	no
B1788			50	no
B1980	Putative glycosyl transferase like protein		234	no
B1995	Putative TonB dependent receptor		139	no
B1997 B0360 B1403 B4272 B2861 B3044	Putative transposase	*yi21_g2*	136	no
B1998	Putative TonB dependent receptor	*yoeA*	168	no
B1999	Putative histone	*sap*	183	no
B2003	Phosphoenolpyruvate carboxylase motif	*yeeT*	73	no
B2006		*yeeW*	64	no
B2085	Putative prokaryotic lipoprotein	*yegR*	125	no
B2116	Putative prokaryotic lipoprotein	*molR_2*	645	no
B2117		*molR_3*	333	
B2118	Predicted ATP/GTP-binding site motifA and Aldehyde dehydrogenases cysteine active site	*yehI*	1210	no
B2119	Putative ATPase	*yehL*	384	no

B2120(2)		*yehM*	759	no
B2121	VWA, von Willebrand factor (vWF) type A domain; found in extracellular proteins.	*yehP*	378	no
B2122		*yehQ*	622	no
B2160	Putative kinase (PfkB family) with DNA binding domains (AsnC, MarR type) carrying Primase, sigma70_r4 and Pint domains	*yeiI*	219	no
B2271	Putative peptidase	*yfbL*	325	no
B2272	Carries and Aminoacyl-transfer RNA synthetases class-II signature 2.	*yfbM*	167	no
B2358	Similar to bacteriophage replication protein O	*yfdO*	122	no
B2363		*yfdT*	101	no
B2373	Putative Thiamine pyrophosphate enzyme, also carries DNA topoisomerase II signature	*yfdU*	564	no
B2376	Putative prokaryotic lipoprotein	*ypdI*	91	no
B2382(1)	Similar to bacterial suface antigen D15	*ypdC*	285	no
B2505	Putative prokaryotic lipoprotein	*yfgH*	172	no
B2545(2)	Putative Zinc binding dehydrogenase	*yphC*	364	no
B2547	Putative ABC transporter	*yphE*	503	no
B2548	Similar to periplasmic binding proteins and carries a sugar binding domain of the LacI family.	*yphF*	327	no
B2549	Carries a tetratricopeptide repeat domain	*yphG*	1124	no
B2767		*ygcO*	98	no
B2769(2)	Similar to electron transfer flavoprotein alpha subunit	*ygcQ*	297	no
B2770	Similar to electron transfer flavoprotein alpha subunit	*ygcR*	261	no
B2870	Putative aspartate/ornithine carbamoyltransferase, carbamoyl-P binding domain	*ygeW*	363	no
B2872	Putative peptidase	*ygeY*	403	no

B2874	Putative amino acid kinase	*yqeA*	310	no
B2876	Putative Tetraacyldisaccharide-1-P 4'-kinase LpxK like.	*yqeC*	235	no
B2877		*ygfJ*	192	no
B2878(1)	Putative oxidoreductase	*ygfK*	1032	no
B2879	Putative proteoglycan. Responsible for decline in cell viability at the beginning of stationary phase	*ssnA*	464	yes
B2880	Predicted FAD binding and CO dehydrogenase domains.	*ygfM*	259	no
B2882	Putative Xanthine, Uracil like permease	*ygfO*	485	no
B2886	Predicted 4Fe4S Ferredoxin like domains	*ygfS*	163	no
B2917	Novel (2R)-methylmalonyl-CoA mutase: converts succinyl-CoA, derived from the tricarboxylic acid cycle, to propionyl-CoA	*sbm*	157	yes
B3013(2)		*yqhG*	309	no
B3036		*ygiA*	86	no
B3063	Putative Sodium:sulfate symporter	*ygjE*	487	no
B3119	Positive regulatory protein for threonine dehydratase, TdcB	*tdcR*	114	yes
B3120	Carries an aldehyde dehydrogenases cysteine active site.	*yhaB*	186	no
B3138	Putative PTS system N-acetylgalactosamine-specific enzyme IIB component 1	*agaB*	158	yes
B3140	Putative PTS system N-acetylgalactosamine-specific enzyme IID component	*agaD*	263	yes
B3491	Putative prokaryotic lipoprotein	*yhiM*	383	no
B3504		*yhiS*	260	no
B3507	Has a predicted LuxR type helix-turn-helix DNA binding domain	*yhiF*	176	no
B3510	Putative prokaryotic lipoprotein	*hdeA*	110	no
B3512	Putative pyruvate kinase with a predicted LuxR	*yhiE*	175	no

	type helix-turn-helix DNA binding domain			
B3517	Glutamate decarboxylase	*gadA*	466	yes
B3665	Putative adenine deaminase	*yicP*	588	no
B3672	*ivlB* operon leader peptide	*ivbL*	32	yes
B3680	Putative prokaryotic lipoprotein with an AraC type helix-turn-helix DNA binding domain	*yidL*	307	no
B3690	Putative monooxygenase	*yidS*	361	no
B3707	Regulatory leader peptide for *tna* operon	*tnaL*	24	yes
B3782	Regulatory leader peptide for *rho* operon	*rhoL*	33	no
B3814			99	no
B3922		*yiiS*	99	no
B3937		*yiiX*	202	no
B3943		*yijE*	312	no
B4109	Putative GTPase	*yjdA*	742	no
B4110		*yjcZ*	281	no
B4126		*yjdI*	76	no
B4127	Putative acetyltransferase	*yjdJ*	90	no
B4309		*yjhS*	326	no
B4313	Recombinase?; regulatory gene for expression of *fimA*	*fimE*	198	yes
B4316	Biosynthesis of fimbriae; periplasmic chaperone for type 1 fimbriae	*fimC*	241	yes
B4320	Membrane-specific adhesin (lectin); major fimbrial subunit; mediates mannose-binding	*fimH*	300	yes
B4334	Belongs to the BcrAD_BadFG, BadF/BadG/BcrA/BcrD ATPase family	*yjiL*	257	no
B4335		*yjiM*	390	no
B4357	Has a GntR type helix-turn-helix binding motif	*yjjM*	268	no

Genes/ORFs in close proximity on the chromosome are marked in grey

3.4. Analysis of E. coli specific ORFs.

The MBGD database reports 133 clusters (above) to be found only in the four *E. coli* genomes and absent in the gamma proteobacterial genomes selected. A large number of these clusters (83) have orthologs in the two *Shigella* genomes on the database. Each cluster may contain more than one ORF, for example, one cluster in table 3.2 contains six ORFs (B1997, B0360, B1403, B4272, B2861, and B3044) which belong to a single protein family of transposases. There are three other clusters which include two ORFs each: 1) b0252 & b2633 2) b0554 & b1556 and 3) b1721 & b4017.

75 of these clusters lie in close proximity to at least one other *E. coli*-specific cluster on the chromosome. The rest are single ORFs whose immediate chromosomal neighbours have orthologs in gamma proteobacterial genomes. 23 clusters contain genes of known function. 88 further clusters have predicted motifs based on amino acid sequence, but the implied functions have not been experimentally confirmed. The average length of amino acid sequences of all 133 clusters is 294.5. 28 proteins were classified as small (<100 amino acids), 81 as medium sized (between 100 & 500 amino acids) and 18 as large (>500 amino acids).

The quick BLASTp facility at the Swissprot database (http://ca.expasy.org/) was used to search for possible homologs of the members of all 133 *E. coli* specific clusters reported above. 67 clusters showed matches in genomes other than those of gamma proteobacteria while 29 showed matches in genomes within gamma proteobacteria. Low sequence identities between orthologs in the gamma proteobacterial genomes may be a reason why the MBGD database reports the 29 clusters as *E. coli* specific. The effect of varying Blastn cut-off values on the number *E. coli* specific genes is discussed in section 3.4.3. 32 clusters did not have any orthologs on the Swissprot database. Some of these species specific genes may of course have orthologs in yet unsequenced genomes.

3.4.1. *E. coli* specific genes since the publication of the K-12 MG1655 genome.

The number of species specific clusters is dependent on three factors. The first of these is the number of genomes used in the comparison and the relation of these to the query genome. The number of *E. coli* specific genes has decreased as the number of sequenced gamma proteobacterial genomes has increased. To understand how *E. coli* specific genes have changed since the completion of the K-12 MG1655 genome the MBGD database was used to find the number of *E. coli* specific genes at different times since 1997. Table 3.2.3 shows the number of *E. coli* specific clusters changed after each new gamma proteobacterial genome was added. The last column (on the right) shows the drop in the number of clusters found to be unspecific when each new genome is added to the analysis.

Table 3.3. The falling numbers of *E. coli* specific genes since 1997:

Name	**Year sequenced**	***E. coli* specific clusters**	**Reduction in specific clusters**
Haemophilus influenzae Rd	Mar-96		
Escherichia coli K-12 MG1655	Sep-97	2464	
Xylella fastidiosa 9a5C	Jul-00	2089	375
Pseudomonas aeruginosa	Aug-00	1365	724
Vibrio cholerae	Aug-00	1135	230
Buchnera sp. APS	Sep-00	1129	6
Escherichia coli EDL933*	Jan-01	1043	86
Escherichia coli 0157:H7*	Feb-01	1025	18
Pasturella multocida PM70	Mar-01	976	49
Yersinia pestis CO92	Oct-01	728	248
Salmonella enterica CT18	Oct-01	366	362

Salmonella typhimurium	Oct-01	323	43
Xanthomonas campestris (ATCC 33913)	May-02	314	9
Xanthomonas axonopodis (306)	May-02	313	1
Buchnera aphidicola (*Schizaphis graminum*)	Jun-02	313	0
Yersinia pestis KIM	Aug-02	312	1
Shigella flexneri 2a 301	Oct-02		
Shewanella oneidensis	Nov-02	300	12
Wiggelsworthia brevipalpsis	Nov-02	300	0
Pseudomonas putida KT2440	Dec-02	281	19
Escherichia coli CFT073*	Dec-02	159	122
Buchnera aphidicola (*Baizongia pistaciae*)	Jan-03	159	0
Xylella fastidiosa Temecula1	Feb-03	159	0
Vibrio parahaemolyticus	Mar-03	149	10
Salmonella enterica Ty2	Apr-03	149	0
Coxiella burnetii	Apr-03	149	0
Shigella flexneri 2a 2457T	May-03		
Candidatus Blochmannia floridanus	Aug-03	149	0
Photorhabdus luminescens	In press	141	8
Haemophilus ducreyi	unpublished	140	1
Pseudomonas syringae pv. tomato	unpublished	137	3
Vibrio vulnificus CMCP6	unpublished	136	1
Vibrio vulnificus YJ016	unpublished	133	3

**The change in* E. coli *specific clusters at these points reflects genes missing from the* E. coli *strain newly included.*

The largest number of shared clusters (with *E. coli*) were found in the following gamma proteobacterial genomes: *P. aeruginosa, X. fastidiosa, S. enterica, Y. pestis, V. cholerae,* the two *E. coli* 0157 genomes combined, *P. multocida* and *S. typhimurium.* Publication of the uropathogenic *E. coli* CFT073 sequence resulted in a large drop in *E. coli* specific

clusters (down 122 unspecific clusters) which takes us to the second factor affecting *E. coli* specific clusters.

3.4.2. Choice of *E. coli* genomes and its effect on the number of *E. coli* specific genes.

The four *E. coli* genomes (the laboratory strain MG1655, two enterohemorrhagic 0157 strains and the uropathogenic CFT073) used in these comparisons are quite diverse from each other sharing only 39.2% of their combined set of proteins (Welch et al, 2002). Addition of each newly sequenced *E. coli* genome to the common *E. coli* cluster set results in a drop in the number of clusters that are specific to *E. coli*. When the two 0157 and the single uropathogenic CFT073 genomes were added to the comparisons the number of specific clusters dropped to 104 and 122 respectively. Considering the two *Shigella* genomes on the MBGD database as non *E. coli* would result in the number of *E. coli* specific clusters falling from 133 to 51.

To find out how the number of *E. coli* specific clusters would change depending on what is considered an *E. coli,* the MBGD database was used to find clusters specific to *E. coli* when *E. coli* is represented by MG1655 alone, MG1655 and CFT073 combined, and MG1655 and the two 0157 genomes combined. Shown below in table 3.4 are the number of clusters which were found to be *E. coli* specific.

Table 3.4. E. coli *specific genes and representative E. coli genomes.*

Representative *E. coli* genome	**Number of shared *E. coli* specific clusters**
MG1655	265
MG1655+ 0157:H7+EDL933	239
MG1655+CFT073	153
MG1655+CFT073+0157:H7+EDL933	133

MG1655+*Shigella flexneri* 2a (301 & 2457T)	117

E. coli specific clusters is highest when *E. coli* is represented by the single K-12 MG1655 genome. The number of specific clusters falls (265 down to 239) when MG1655 and the two 0157 genomes are combined to form a single search set. However *E. coli* specific clusters appear much smaller (265 down to 153) when MG1655 and the uropathogenic CFT073 are combined to form a representative *E. coli* genome. This perhaps reflects the differences in genome content and lifestyles of the two *E. coli* strains. When the K-12 and two *Shigella* genomes are combined as a search set the number of shared *E. coli* specific clusters fell even further down to 117. This was unexpected since the two *Shigella* genomes are also very closely related to the K-12 MG1655 genome sharing approximately 3.9 megabases of the chromosomal 'backbone' sequence.

3.4.3. BLASTn cut-off values and *E. coli* specific genes.

The third factor that affects the number of *E. coli* specific clusters although to a lesser extent than the two factors above is the chosen BLASTn cutoff value. BLAST cutoff values from $p=10^{-2}$ to $p=10^{-9}$ were used to find the numbers of *E. coli* specific clusters. Other parameters such as the query and reference genomes were the same as for table 3.2. The number of *E. coli* specific clusters rises from 133 at $p=10^{-2}$ to 165 at $p=10^{-9}$. This slight increase in the number of species specific sequences on decreasing the BLAST probability cut-off values is similar to those reported by the makers of the Neurogadgets web resource during their search for species specific sequences (Charlebois et al, 2003)

3.5: Functional analysis of E. coli *specific genes*

3.5.1: Deletion of *E. coli* specific genes

The function of *E. coli*-specific clusters were investigated by deleting the corresponding genes from the chromosome of *E. coli* K-12 MG1655. The genes were deleted using the modified pKO3 deletion method in which genes are replaced with a *lacZ aph* reporter cassette. The reporter cassette also contains a *lac* promoter (*plac*) at the distal end to minimise the downstream effects of integrating the cassette. The *lacZ* gene was used to collect expression data from the native promoters of deleted genes and *aph* was a useful selectable marker used in the deletion procedure. The cassette is flanked by FRT sites which are recognised by FLP recombinase which can be used to 'flip' the cassette out from the deletion leaving behind an in-frame deletion scar (Merlin et al. 2002). The deletion process, primers and condition used are described in Materials and Methods 2.2, 2.14 and 2.15.

A total of 49 ORFs were seperately deleted from the chromosome of *E. coli* K12 MG1655. Four mutants have multiple deletions of adjacent genes namely *ΔyahL, M, ΔygjM, N, ΔyjdI, J, K* and *ΔydhR, S, T, U, X, W, V, Y* and *Z* were deleted together in the same experiment. These genes are marked in grey boxes in table 3.5 below.

One family of ORFs named the *yhcN* family includes the ORFs *yhcN, ybiJ, ybiM, ycfR, ydgH, yjfY, yahO, yjfN, yjfO* and *ykgI* (highlighted with a '$^{\Theta}$' symbol in table 3.5). There is no function associated with any member of this gene family, but all contain a hypothetical signal sequence at the N terminal end of the predicted protein suggesting that they may be exported to the periplasm. All members of the family also show a conserved C-terminal domain and all except *ykgI* show another conserved N-terminal domain both of unknown function. These two domains are collectively termed *yhcN* like repeat. The largest member of the family, *ydgH*, is composed of three *yhcN* like repeats. *YjfO* has been classified as a putative lipoprotein (Rudd, 1998).

All members of this family, except for *yjfN* and *yjfO,* were deleted individually from the parent MG1655 Δ*lacZ* strain. All deletions were subsequently moved into a single strain using P1 transduction. This involved 'flipping' the reporter cassette using the FLP

recombinase on plasmid pCP20 to leave a null in-frame deletion in place of the wild type sequence and then transducing the next member into the combination mutant. The last two members of the family (*yjfN* and *yjfO*) remain to be transduced into the combined deletion strain by Prof. Kenneth Rudd and the final mutant will be used in phenotypic tests on the Biolog system (Bochner, 2003) to detect any differences in growth compared to the parent strain. The rationale behind this deletion combination was that although deletions of individual members of the family may not produce a detectable phenotype, deleting all members in a single strain may produce a phenotypic effect.

The table below also lists the lengths and positions of genes deleted, the percent of the ORF sequence deleted and any known/predicted motifs or experimental data. The column labeled 'Orthologs in other genomes' contains information on any orthologs of the target genes found in gamma-proteobacterial or other genomes sequenced after the target selection had been made.

Table 3.5: List of ORFs deleted, arranged according to ascending map position.

Gene /ORF name	**Gene length (bp)**	**Percent deleted**	**Map position**	**Orthologs in other genomes (%=percent identity in MBGD as of 30.10.03).**	**Predicted motifs/function and published experimental work**
ykgI$^{\Theta}$	252	63.09	6.85	*Salmonella*	*yhcN* family member, potential signal sequence.
yahL	816		7.4	*E. coli* only	No predicted domains.
yahM	276	96.94	7.43	*E. coli* only	No predicted domains.

yahO$^{\Theta}$	276	65.21	7.45	*Salmonella*	*yhcN* family member, potential signal sequence. Reported under *rpoS* control in *Salmonella* (Ibanez-Ruiz et al 2000)
ybhC	1284	94.93	17.36	*Salmonella, Shigella, Yersinia*(25%), *Xanthomonas* (31%).	Putative lipid anchor and pectinestrase domains. Protein isolated on 2D gel (Lai et al, 2004).
ybiJ$^{\Theta}$	261	74.71	18.04	*Salmonella, Shigella, Yersinia* (44%).	*yhcN* family member, potential signal sequence.
ybiM$^{\Theta}$	405	48.14	18.13	*Shigella.*	*yhcN* family member, potential signal sequence.
yccV	369	74.59	22.15	*Salmonella, Shigella, Wigglesworthia, Yersinia* (75%), *Candidatus. floridanus.*	Function unknown. Upregulated during heat shock (Richmond et al, 1999).
yceP	255	63.09	24.14	*Salmonella, Shigella.*	Function unknown. Upregulated during heat shock (Richmond et al, 1999) and oxidative stress (Zheng et al, 2001).
ycfR$^{\Theta}$	258	72.94	25.18	*Salmonella, Shigella.*	Function unknown. Reported to be upregulated during heat shock (Richmond et al, 1999).

ycjT	2268	97.75	29.66	*Shigella.*	Probable glycosyl hydrolase family: family of proteins that cleave the glycosidic bonds between two carbohydrates/non-carbohydrates. Includes vacuolar acid hydrolase & maltose phosphorylase. Present in the *hslE-H* heat shock loci.§
yncE	1062	96.04	32.79	*Salmonella, P. putida* (21%), *Xanthomonas* (23%)	Predicted bacterial Pyrrolo-quinoline quinone (PQQ): is a redox coenzyme, which serves as a cofactor for a number of enzymes (quinoproteins) and particularly for some bacterial dehydrogenases. A number of bacterial quinoproteins belong to this family.§
ydeK	3978	98.94	34.32	*E. coli, Shigella.*	Hypothetical lipoprotein.
ydgH$^{\Theta}$	945	93.33	36.14	*Salmonella, Shigella, Yersinia* (61%)	*yhcN* family member, potential signal sequence.
ydhR	306	98.68	37.61	*Salmonella, Shigella, Yersinia* (50%), *V. parahaemolyticus* (60%)	Potential signal sequence.
ydhS	1605		37.62	*E. coli* only.	No predicted domains.
ydhT	813		37.65	*Shigella.*	No predicted domains.
ydhU	786		37.67	*Salmonella, Shigella.*	Predicted domain: Nickel-dependent hydrogenase b-type cytochrome subunit. Similar to *phsC* in Salmonella. Functions as a membrane anchoring protein for the Phs system

					oxidoreductase that produces hydrogen sulfide from thiosulfate (potential).§
ydhX	720		37.69	*E. coli*, *Shigella*, *Salmonella.*	Similarity to *E. coli nrfC*. The iron-sulfur centers are similar to those of 'bacterial-type' 4FE-4S ferredoxins.§
ydhW	648		37.7	*E. coli* only.	No domains predicted.
ydhV	2103		37.72	*Shigella*, *Pyrococcus abyssi**.	Belongs to a family of aldehyde ferredoxin oxidoreductases found mainly in archaea. Enzymes of the aldehyde ferredoxin oxidoreductase (AOR) family contain a tungsten cofactor and a 4Fe4S cluster and catalyse the interconversion of aldehydes and carboxylic acids.§
ydhY	627		37.76	*Shigella.*	similar to PhsB in *Salmonella typhi*. The iron-sulfur centers are similar to those of 'bacterial-type' 4FE-4S ferredoxins.§
ydhZ	210		37.79	*Salmonella*, *Shigella.*	No predicted domains.
yedJ	696	91.81	43.77	*Salmonella*, *P. putida* (40%), *V. parahaemolyticus* (40%)	Probable metal dependent phosphohydrolase. Family includes *relA* and *spoT*. Probable functions may include nucleic acid metabolism and signal transduction.§
yegI	1947	96.76	46.28	*Shigella*, *Pseudomonas syringeae* (28%).	Two hypothetical motifs: 1) helix-hairpin-helix motif (non-specific DNA binding protein). 2) protein kinase motif. §

yegR	378	82.54	46.69	*E. coli* only	Expression controlled by *evgA*. Deletion does not affect acid survival (Masuda and Church, 2002).
yeiN	939	96.48	48.62	*Shigella*	71% similar to protein IdgA involved in pigment biosynthesis in *Erwinia chrysanthemi*- Indigoidine is a blue pigment synthesised *by Erwinia chrysanthemi* implicated in pathogenicity and protection from oxidative stress. §
yfbL	978	85.17	51.39	*Shewanella* (24%), *V. cholerae* (31%), *Xanthomonas* (23%)	Predicted peptidase M28: Metalloproteases are the most diverse of the four main types of protease, with more than 30 families identified to date. In these enzymes, a divalent cation, usually zinc, activates the water molecule. The metal ion is held in place by amino acid ligands, usually three in number. The known metal ligands are His, Glu, Asp or Lys and at least one other residue is required for catalysis, which may play an electrophillic role. Of the known metalloproteases, around half contain an HEXXH motif, which has been shown in crystallographic studies to form part of the metal-binding site.§
yfjP	870	95.51	59.62	*Shigella*, *V. cholerae* (23%).	similar to *yeeP* and *ykfA*. Three potential GTP binding domains. Plus two domains of unknown function.§
ypjC	483	83.22	59.98	*E. coli* only.	No predicted domains.

ygaQ	333	79.27	60.02	*E. coli* only.	No predicted domains.
yqhG	930	93.22	68.02	*Shigella.*	Potential signal sequence.§
ygiN	315	82.85	68.36	*Salmonella*, *Shigella.*	Predicted antibiotic biosynthesis monooxygenase-ABM (involved in production of antibiotics in *Streptomyces coelicolor*).§ Isolated as a low abundance protein (Fountoulakis et al, 1999).
ygjMN	417, 315	96.7	69.95	*Salmonella*, *Shigella*, *Shewanella* (43%), *C. burnetti* (41%)	Predicted H-T-H motif. Belongs to a large family of DNA binding proteins in bacteria for example a bacterial plasmid copy control protein, bacterial methylases, various bacteriophage transcription control proteins and a vegetative specific protein from slime mould.§
yraQ	1041	96.82	71.02	*Shewanella*, *Shigella.*	Belongs to a family of predicted permeases of unknown specificity.§
yhcN$^{\Theta}$	315	79.04	72.93	*Salmonella*, *Shigella*	*yhcN* family member, potential signal sequence
yhiM	1152	96.87	78.3	*E. coli* only	No predicted domains. 62% similar to a hypothetical protein in *C. perfringens.*
hdeB	339	69.61	78.75	*Salmonella*, *Shigella*, *Yersinia*(43%)	similar to *hdeA*. Protein structure of both A and B known. Probable periplasmic location.
hdeA	333	77.47	78.76	*Shigella*	Potential periplasmic protein. Reported H-Ns dependent expression (Yoshida et al, 1993). *HdeA* and *hdeB* mutants in *E. coli* are not sensitive to acid shock (Tucker et al, 2002).

yigE	486	92.59	86.18	*Salmonella, Shigella, R. meliloti*, Agrobacterium*.*	No predicted motifs
yihR	927		87.67	*Salmonella, Shigella, C. burnetti* (23%), *P. putida* (24%)	Predicted motif: Aldose 1-epimerase (EC: 5.1.3.3) (mutarotase) is the enzyme responsible for the anomeric interconversion of D-glucose and other aldoses between their alpha and beta forms.§
htrC	540	84.44	90.26	*E. coli* only	Reportedly essential for heat shock survival (Raina and Georgopolous, 1990) .
yjdA	2229	97.84	93.22	*Shigella.*	Predicted H-T-H motif.§ Lies downstream of the *phn* operon but is not involved in methylphosphonate utilization. ATP binding domain might be required for growth in LB (Badarinarayana et al 2001).
yjdI		95.66	93.75	*E. coli* only	No predicted domains.
yjdJ			93.76	*E. coli* only	Predicted: GCN5-related N-acetyltransferase: Histone acetylation is carried out by a class of enzymes known as histone acetyltransferases (HATs), which catalyze the transfer of an acetyl group from acetyl-CoA to the lysine E -amino groups on the N-terminal tails of histones. Early indication that HATs were involved in transcription came from the observation that in actively transcribed regions of chromatin,

					histones tend to be hyperacetylated, whereas in transcriptionally silent regions histones are hypoacetylated.§
yjdK			93.77	*E. coli* only	No predicted domains.
yjfY$^{\Theta}$	276	81.52	95.32	*Salmonella, Shigella.*	*yhcN* family member, potential signal sequence
yjiW	399	87.97	98.66	*Salmonella, Shigella Yersinia* (42%)	No predicted domains. LexA regulated, may be involved in host restriction modification system. Lies between *hsdS* and *mcrB* (Fernandez et al, . 2000).

**-These genomes do not belong to the gamma proteobacteria family.*

§-Motif and functional predictions obtained from SwissProt.

$^{\Theta}$*-Members of the* yhcN *gene family.*

Gene deletion is an important way to test the essentiality of gene function. The deletion procedure is dependent on two homologous recombination events. The first event results in the incorporation of the deletion plasmid into the chromosome. The second recombination event results in an excision of the plasmid from the chromosome leaving the deleted gene::reporter construct on the chromosome followed by loss of the gene on the plasmid. This last step proves unsuccessful if the gene being deleted encodes a function that is essential to the cell. All the genes selected for deletion in this study were successfully deleted and since all the deletion work was carried out in rich medium (LB

broth or LB agar) this means that none of the genes deleted is essential to the cell in this growth medium.

3.5.2: Phenotypic tests

The first set of phenotypic tests were carried out to detect any differences in growth rates and gene expression of the mutants during growth in LB broth at 37 °C. The second set of tests examined growth on solid media at different temperatures and without oxygen and on different media whose components affect growth of the cells. This set of tests was carried out to detect conditional phenotypes that may have resulted from gene deletion.

3.5.2.1: Growth and gene expression in LB broth at 37°C.

All the genes deleted except for *htrC, hdeA* and *hdeB* are *'y'* or predicted, hypothetical genes; therefore it was of interest to see if and how these genes were expressed during growth. Overnight cultures of all strains were inoculated into fresh LB broth and after two subsequent dilutions in mid-logarithmic phase were allowed to grow until they reached the stationary phase. Growth curves of all mutants and of the parent MG1655 Δ*lacZ* (EDCM367) strain were then compared. Gene expression was measured in all phases of growth by assaying the levels of β-galactosidase produced from the *lacZ* reporter cassette under the control of any native promoters. Gene replacement, measurement of growth and β-galactosidase assays were carried out as detailed in Materials and Methods (sections 2.14, 2.15 and 2.16).

All mutants grew similarly to the parent strain showing that none of the mutations results in differences in cell-doubling times in the conditions tested. Gene expression levels and patterns however different amongst genes. Gene expression was measured by assaying the levels of ß-galactosidase produced from the transcriptional fusion *lacZ* under the control of the unaltered native promoter of each gene target.

Expression values are expressed as Miller units and total expression units. Miller units are calculated as detailed in section 2.16 in chapter 2.

All growth curves and the corresponding ß-galactosidase assay values are shown in figures 3.4 to 3.10. Of the 37 ß-galactosidase assays carried out, ORFs *ygjMN, yhcN, hdeB, yceP, yahO, yhiM, hdeA, ydgH* and *yccV* in order of decreasing expression showed high levels of expression, typically between 500 to 2500 Miller units. Several

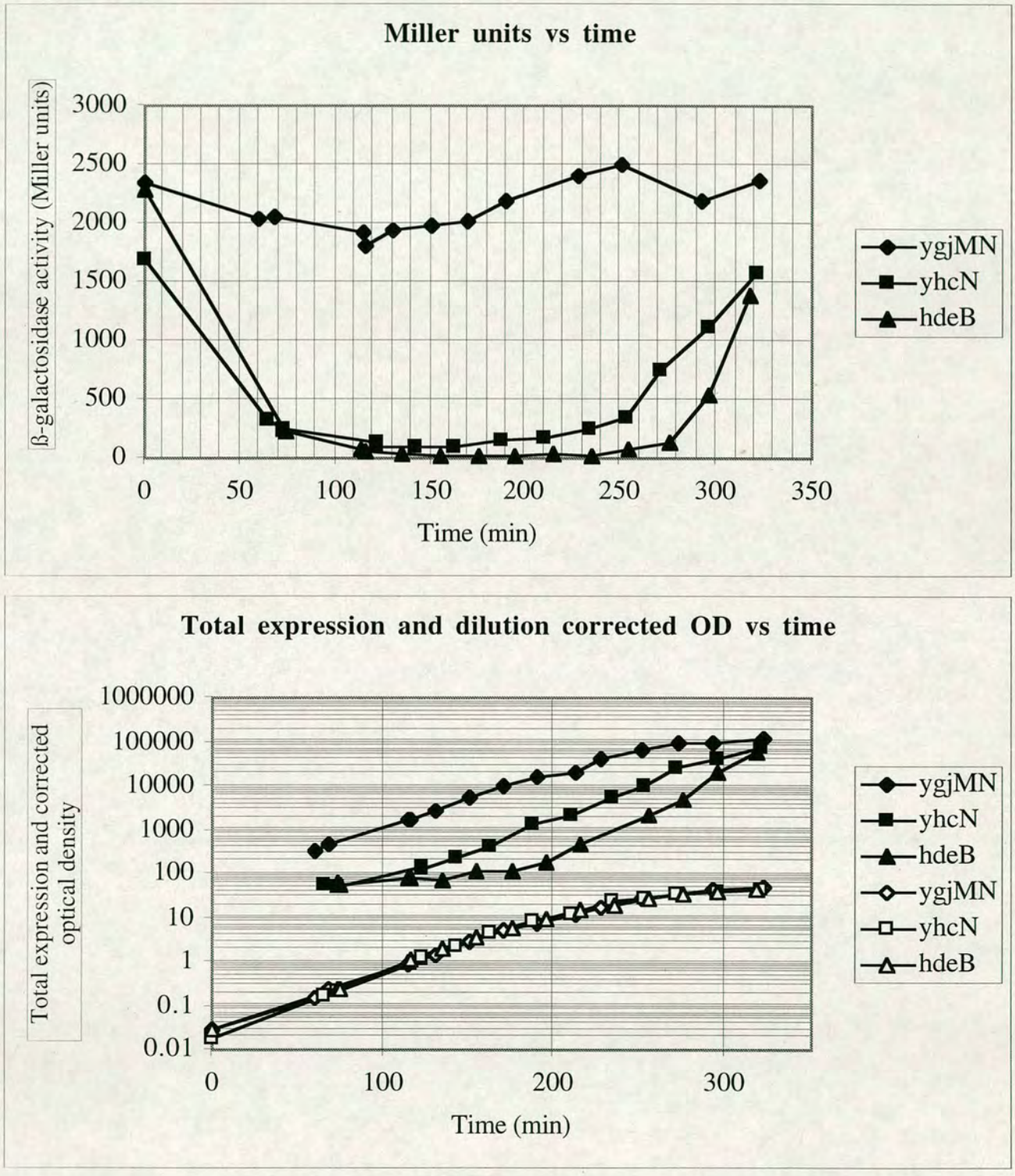


Figure 3.4: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) ygjMN, yhcN *and* hdeB

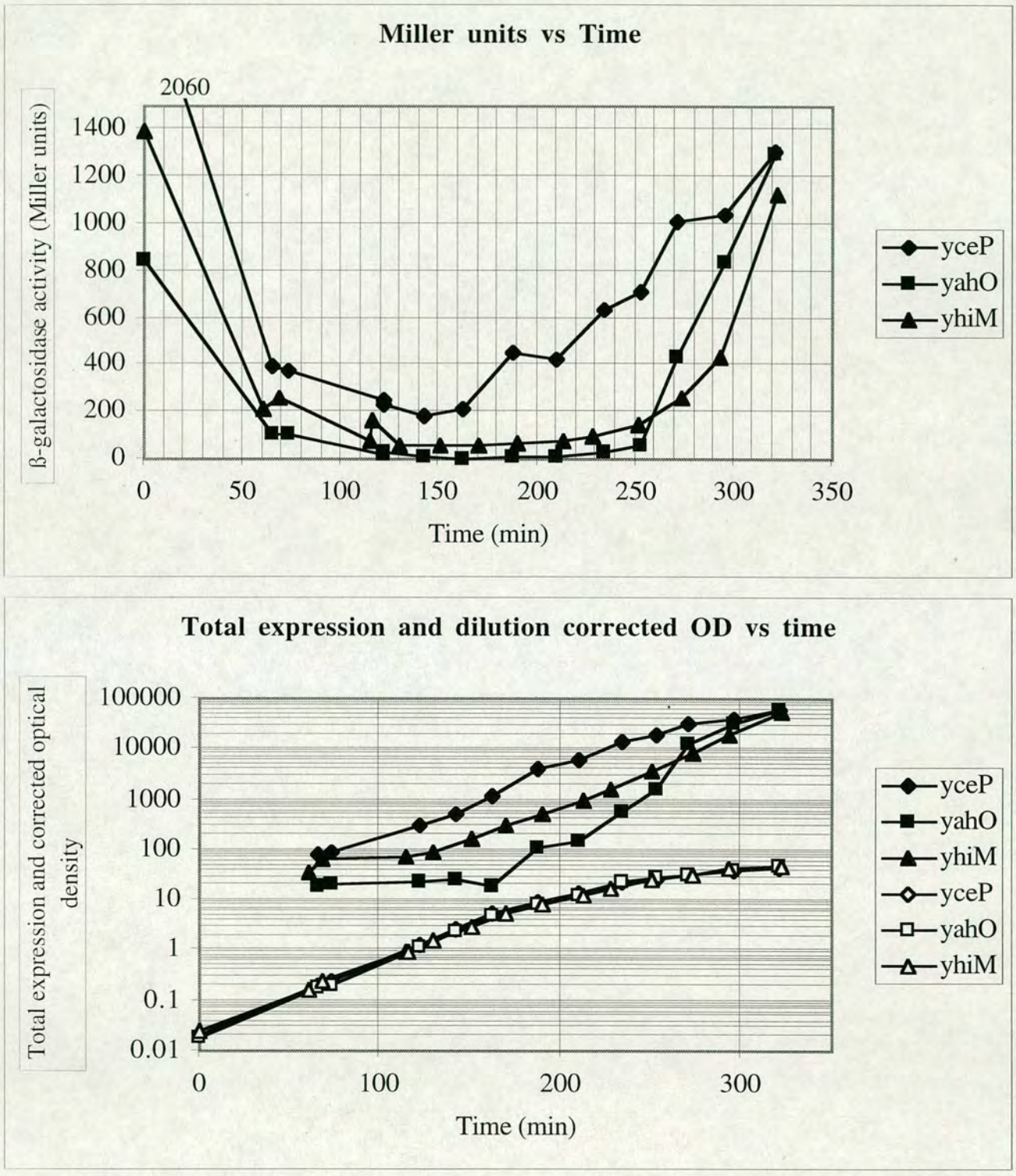
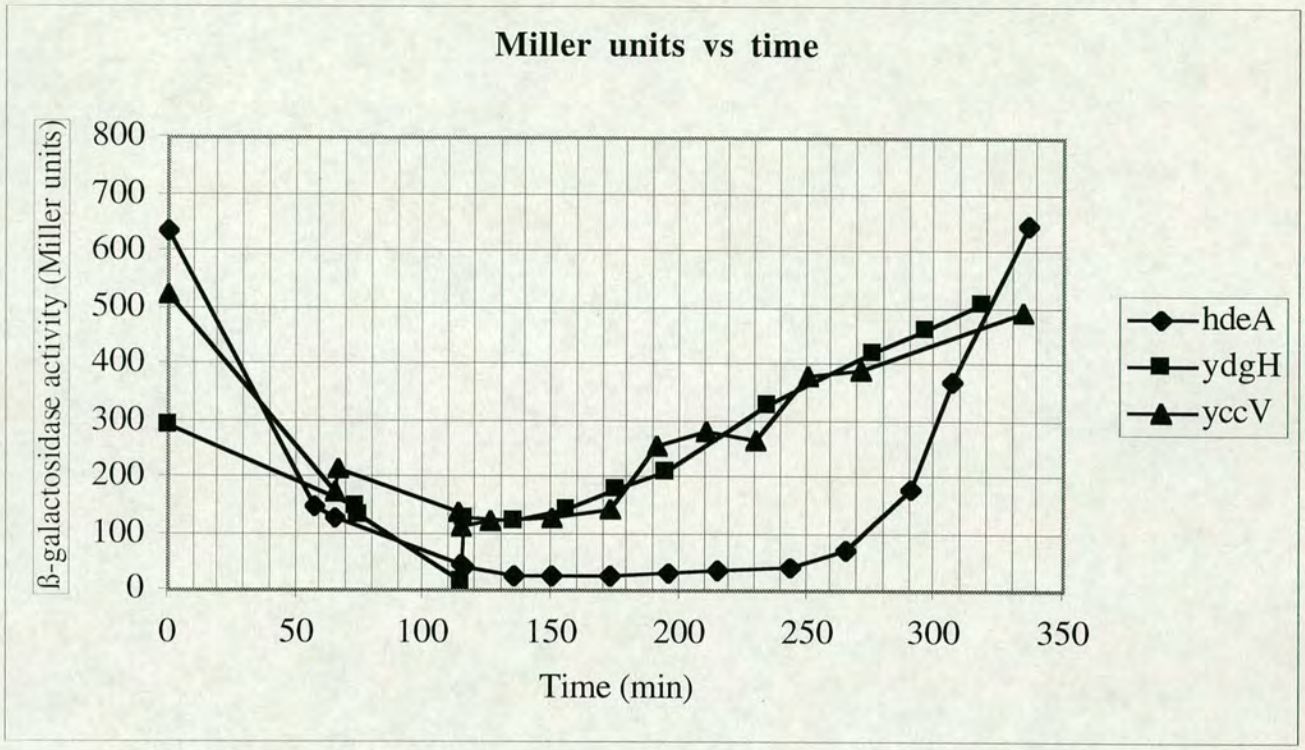


Figure 3.5: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) yceP, yahO *and* yhiM.



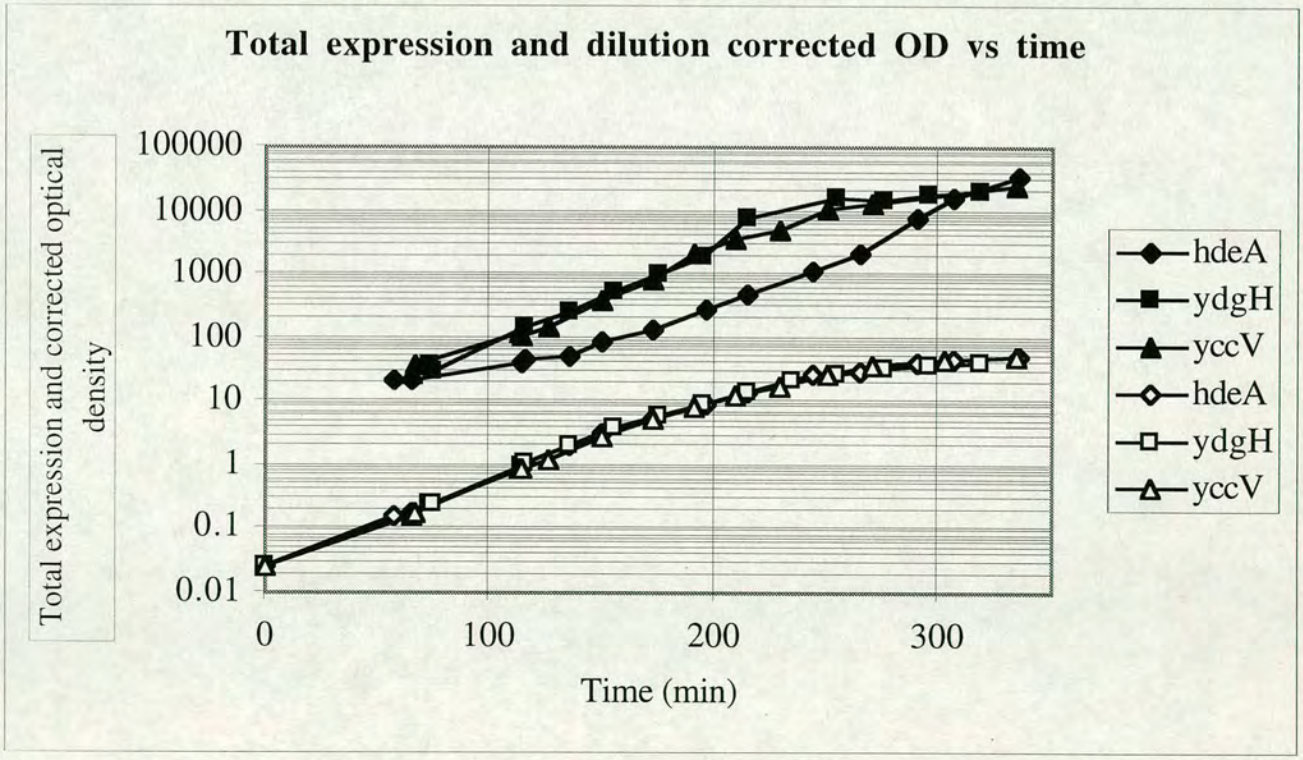


Figure 3.6: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) hdeA, ydgH *and* yccV.

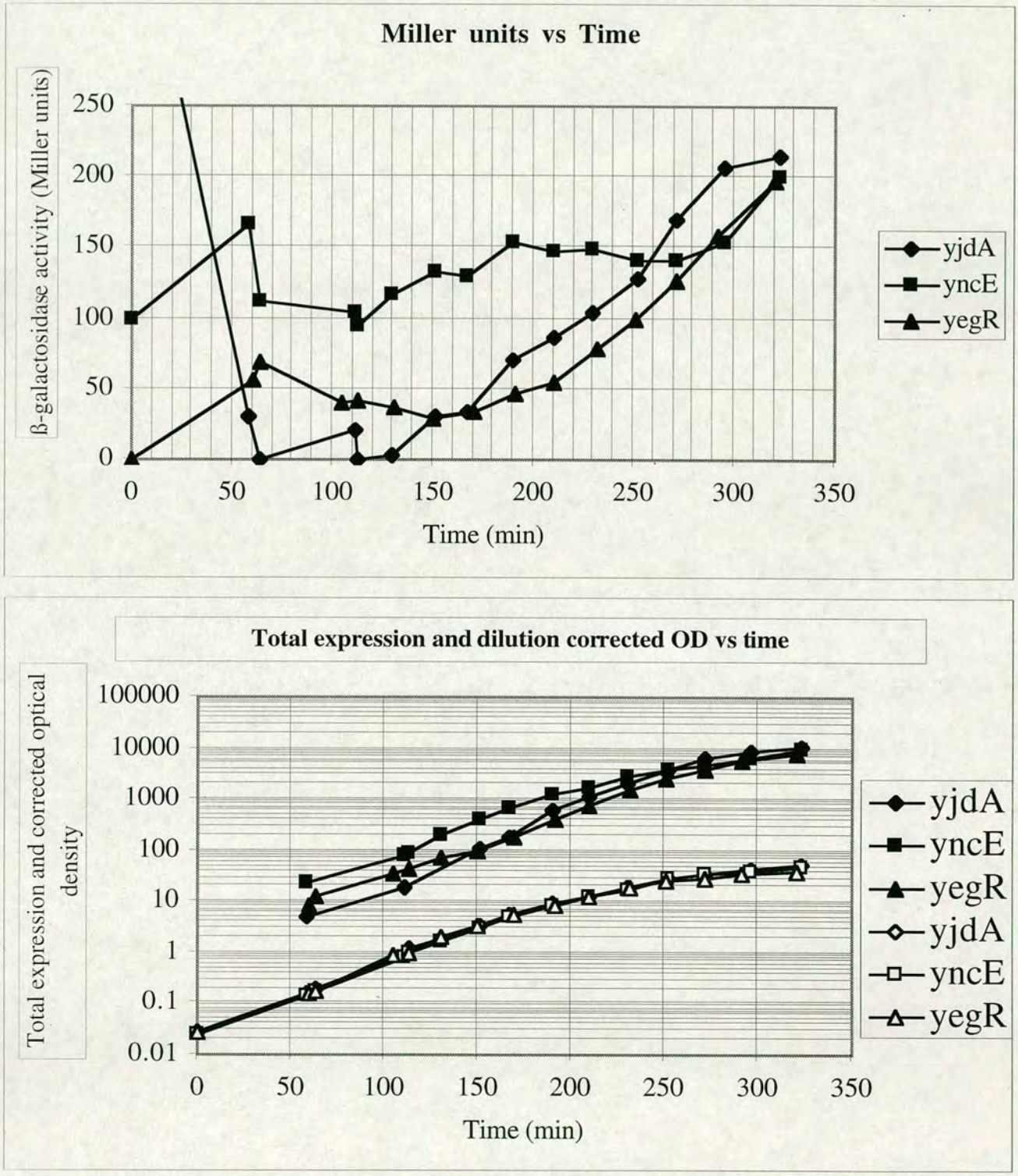


Figure 3.7: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) yjdA, yncE *and* yegR.

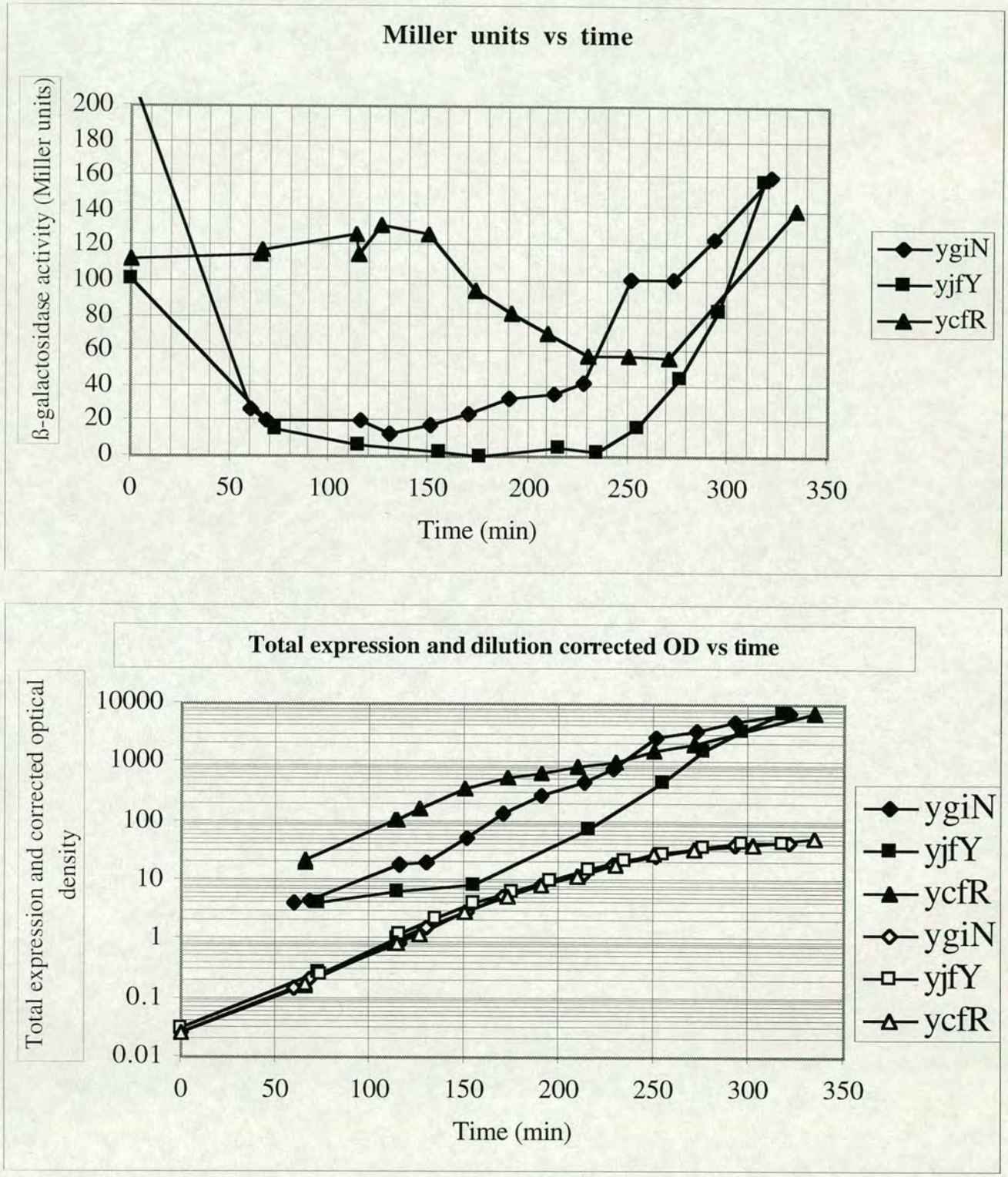


Figure 3.8: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) ygiN, yjfY *and* ycfR.

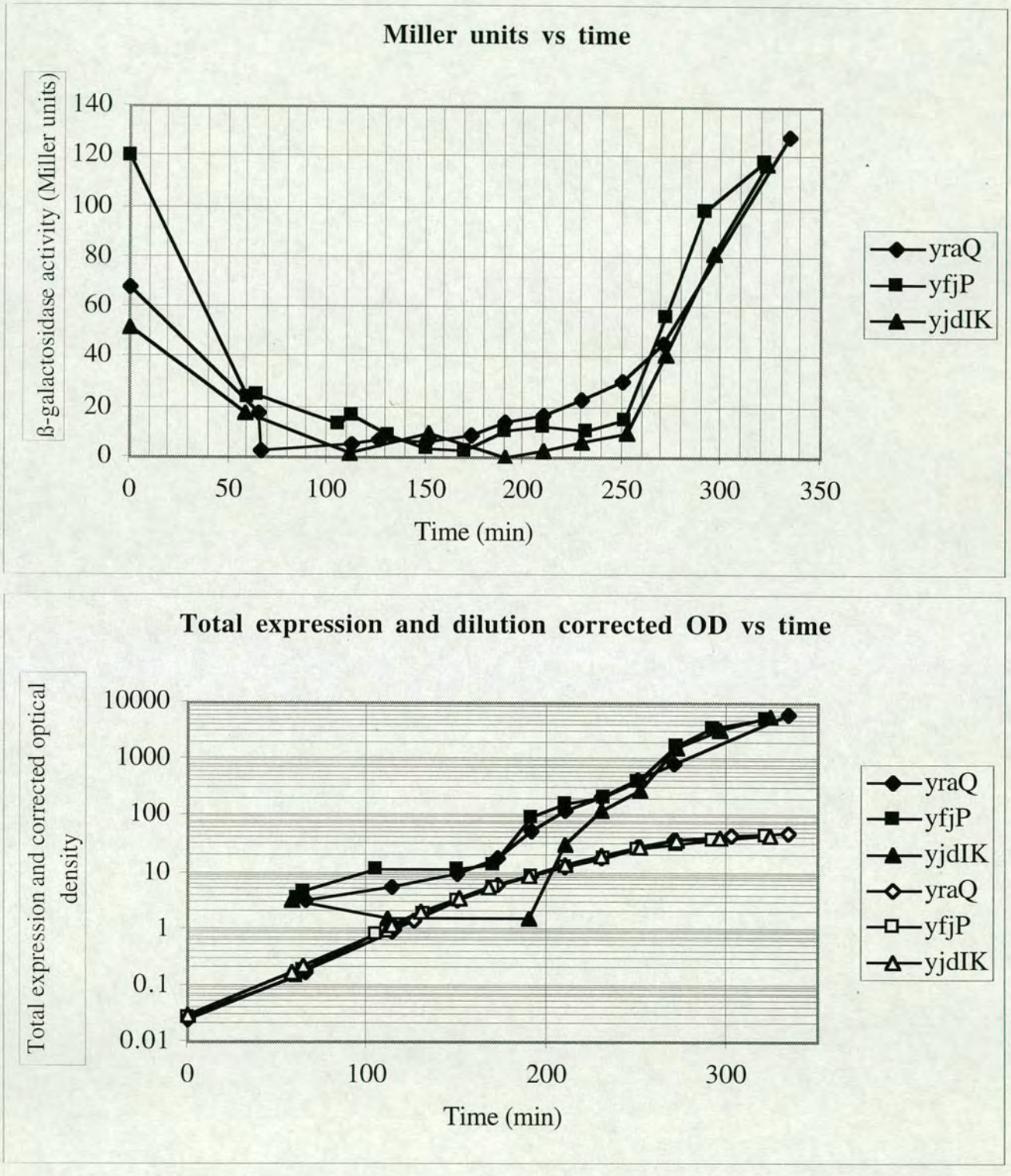


Figure 3.9: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) yraQ, yfjP *and* yjdI-K.

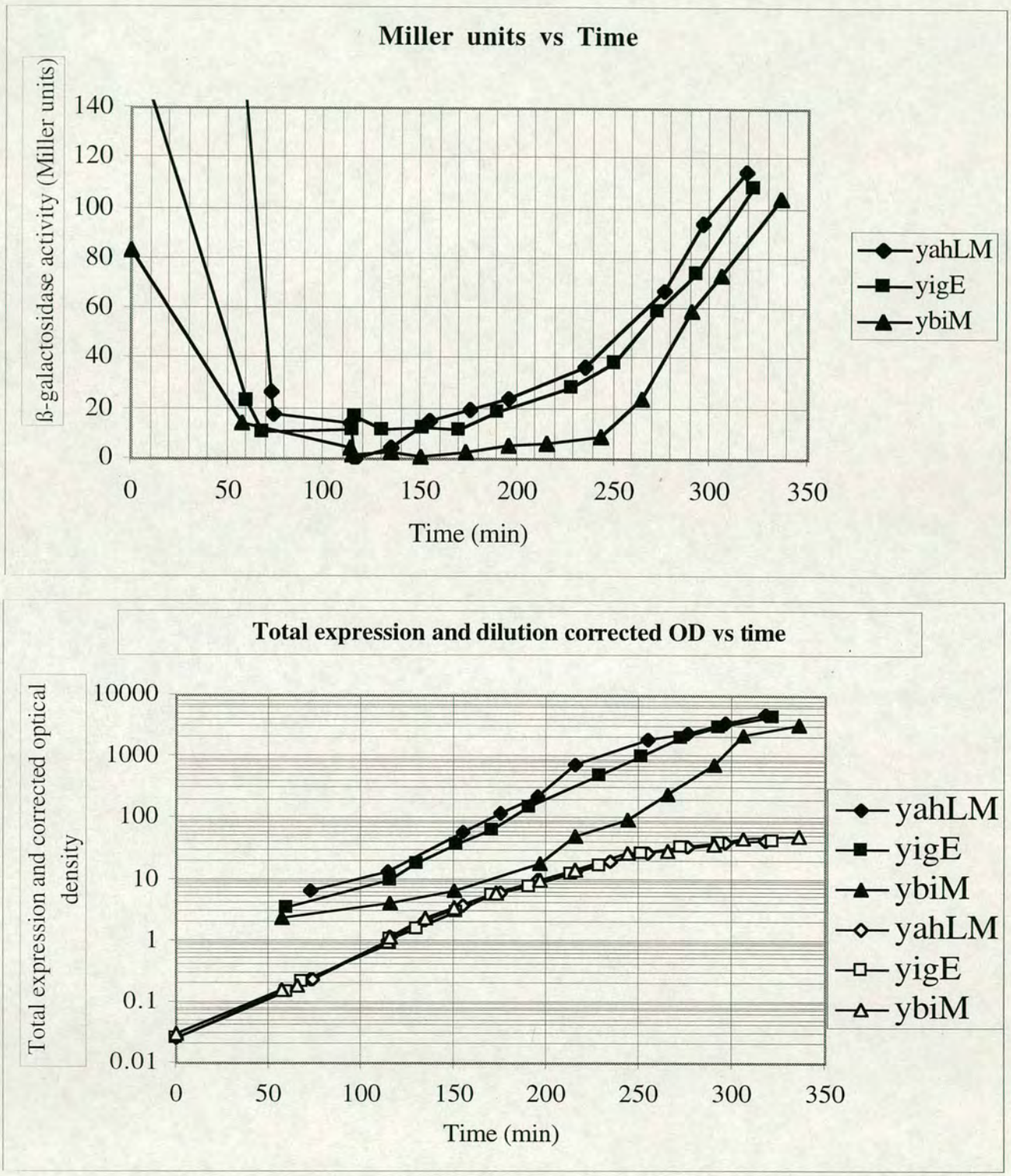


Figure 3.10: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) yahLM, yigE *and* ybiM.

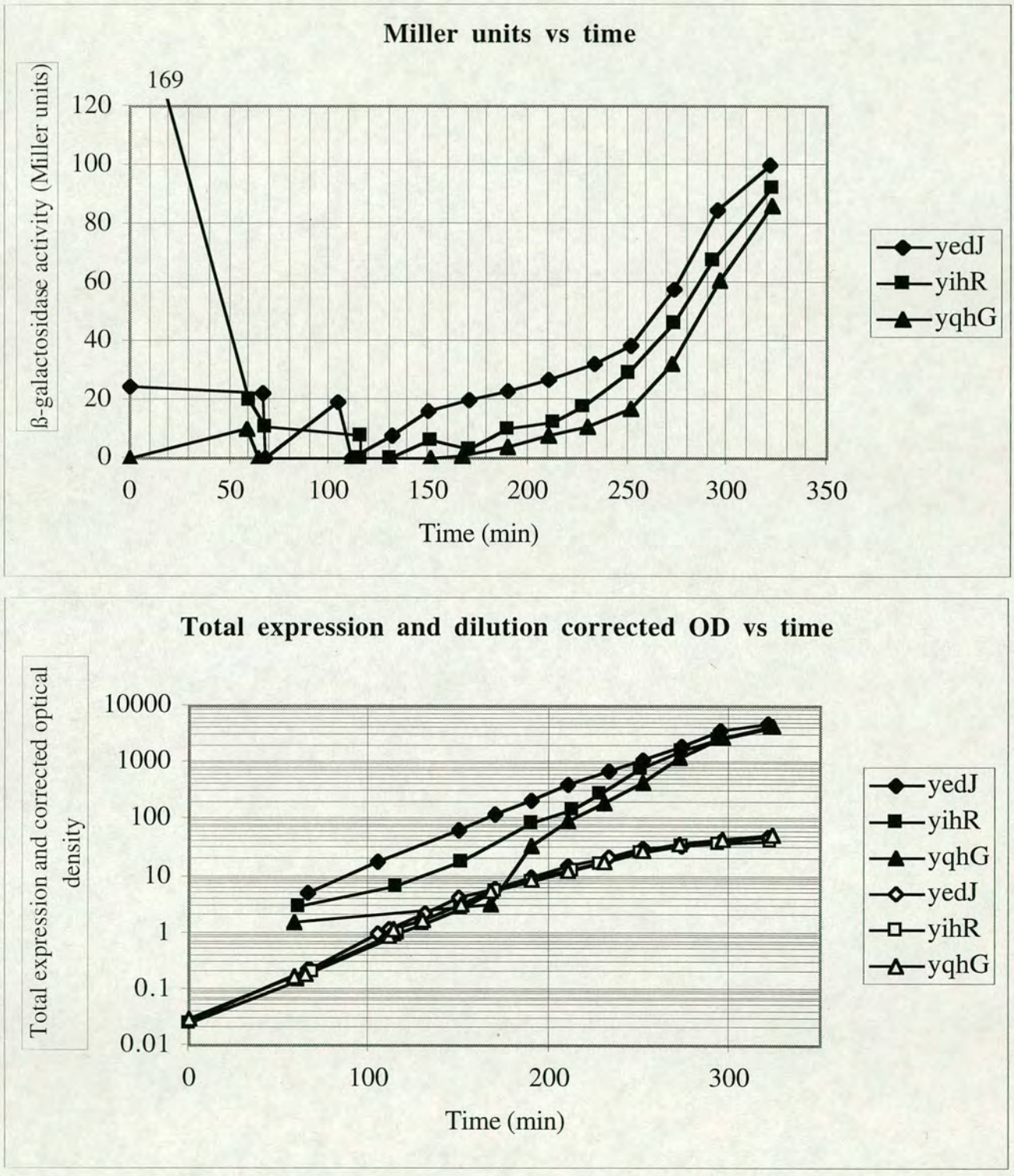


Figure 3.11: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) yedJ, yihR *and* yqhG.

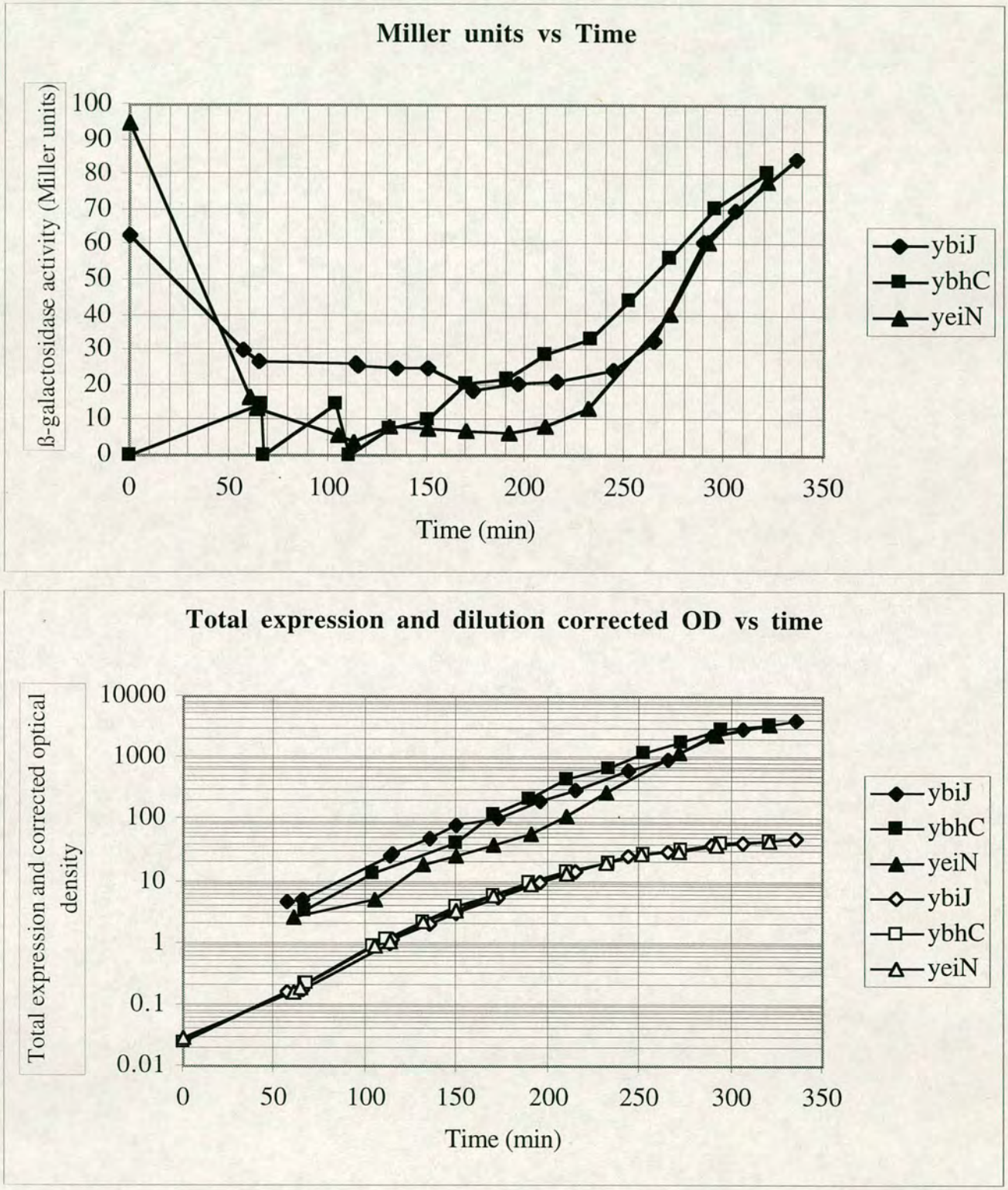


Figure 3.12: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) ybiJ, ybhC *and* yeiN.

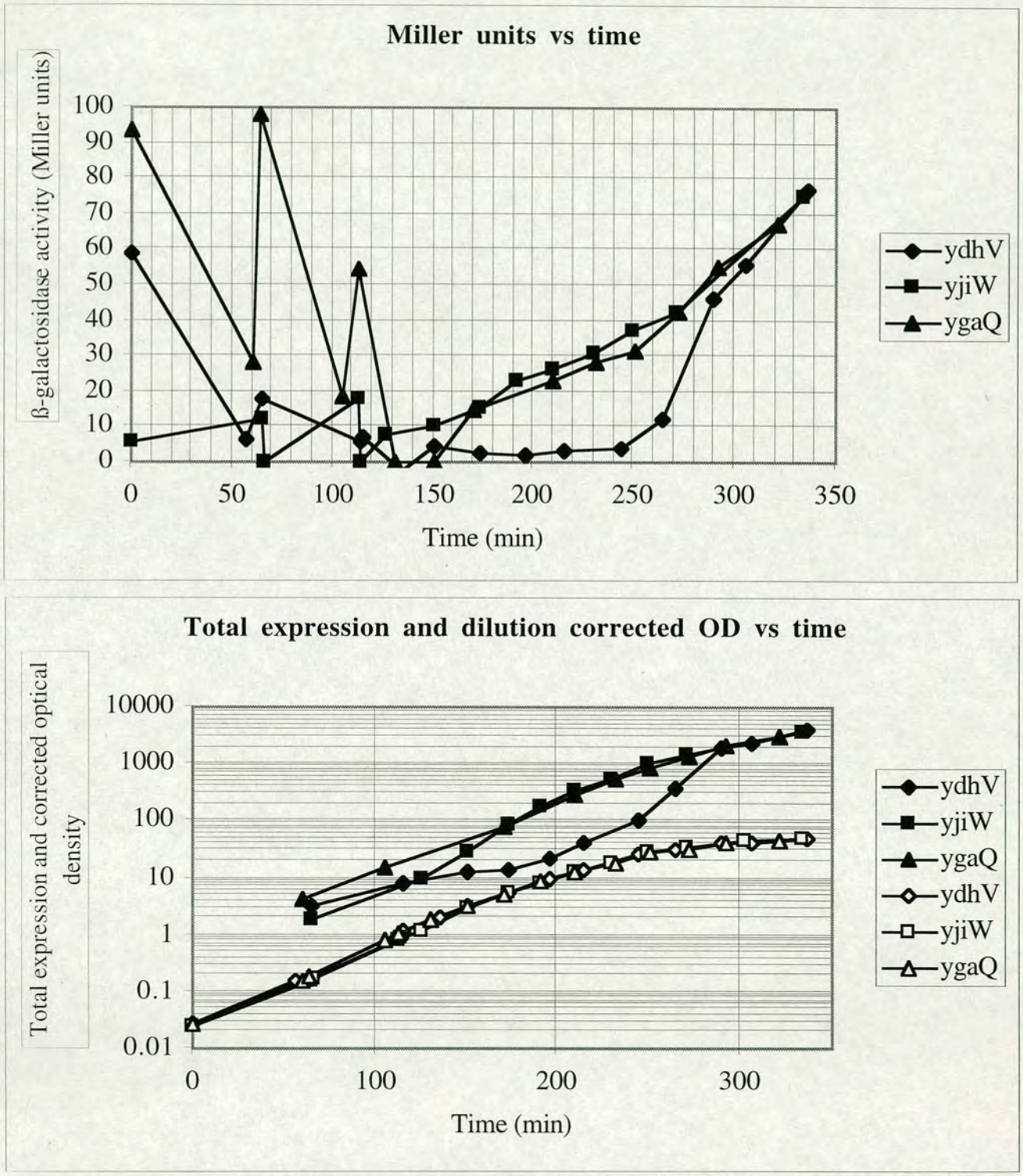


Figure 3.13: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) ydhV, yjiW *and* ygaQ.

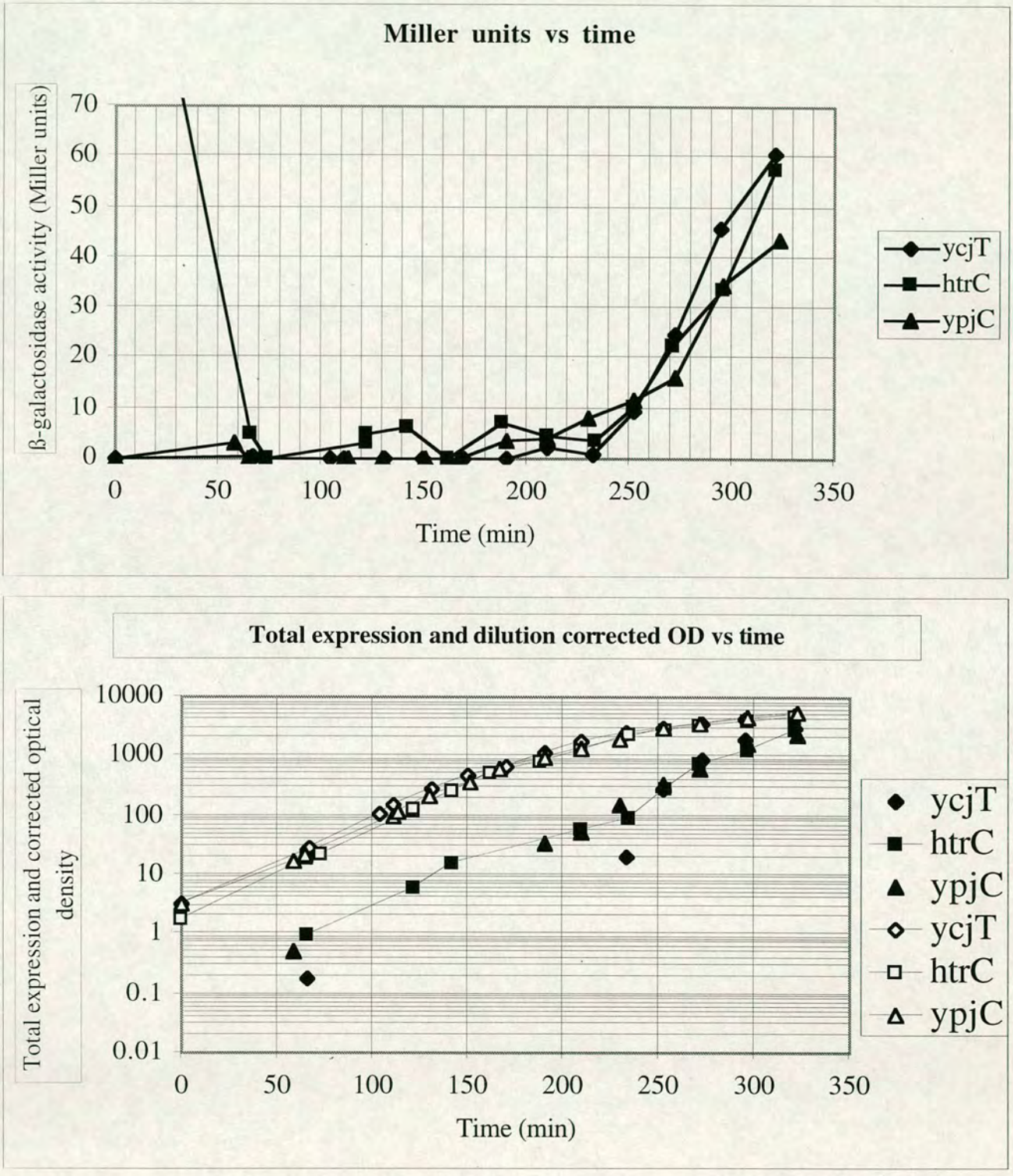


Figure 3.14: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) ycjT, htrC *and* ypjC.

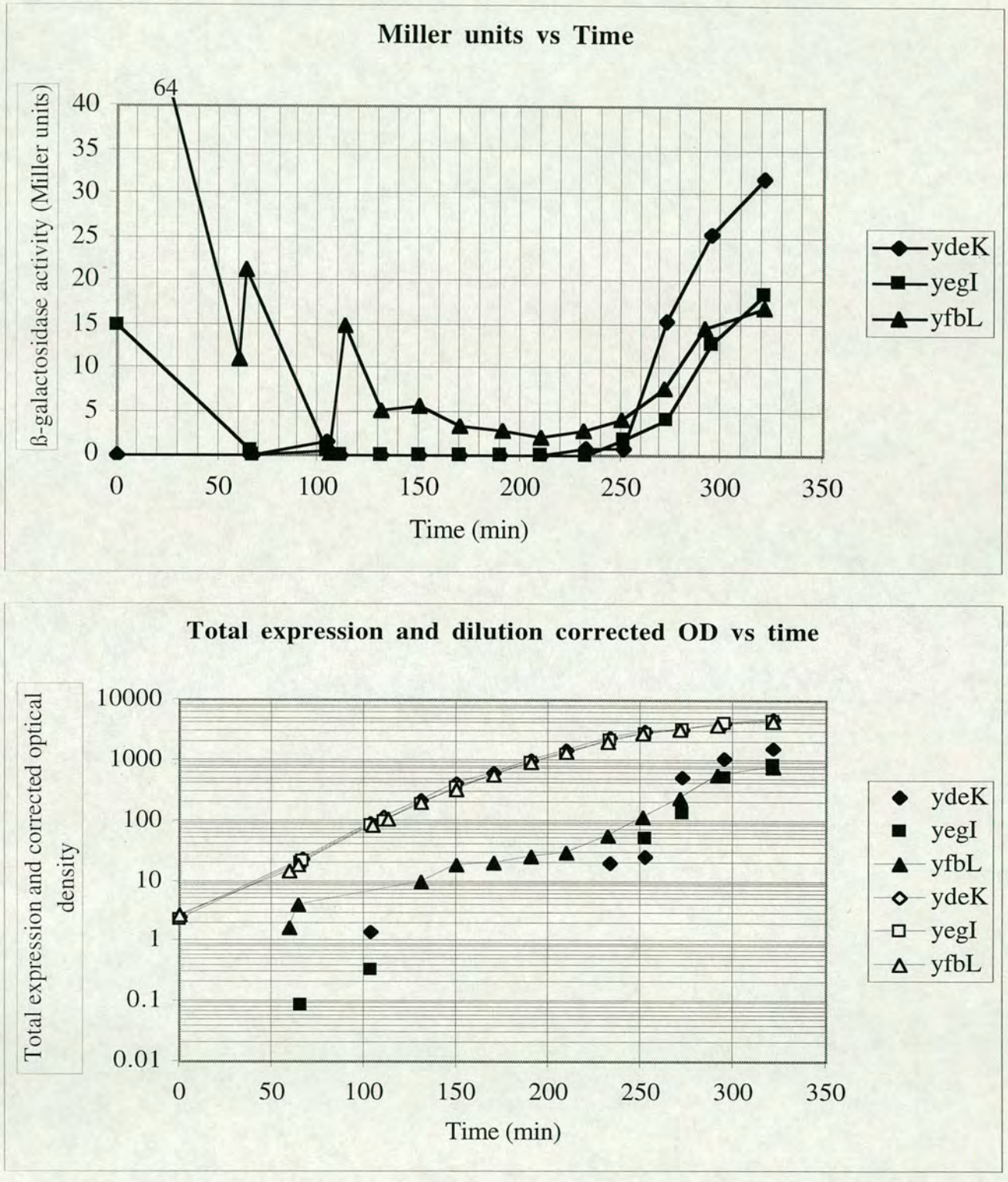


Figure 3.15: Growth curve (open symbols), expression in Miller units and total expression of ORFs (closed symbols) ydeK, yegI *and* yfbL.

other genes showed intermediate levels of expression, between 250 to 50 Miller units. These are (in order of decreasing expression), *yjdA, yncE, yegR, ygiN, yjfY, ycfR, yraQ, yfjP, yjdI-K, yahLM, yigE, ybiM, yedJ, yihR, yqhG, ybiJ, ybhC, yeiN, ydhV, yjiW, ygaQ, ycjT* and *htrC*. ORFs *ypjC, ydeK, yegI* and *yfbL* showed levels of expression below 50 Miller units even at the end of growth when Miller units of expression tend to increase. The levels of gene expression used to categorise ORFs into low, intermediate and high expression categories were based on the last time point measured in the growth curves shown in figures 3.4 to 3.15, at the point at which sampling was stopped. Since many ORFs show an increase in expression during entry into stationary phase the maximal expression levels may be higher than are seen here.

Along with differences in levels of gene expression many ORFs also show differential patterns of expression in various phases of growth. To appreciate how expression patterns vary over time and during different phases of growth, expression values were plotted in two different ways in figures 3.4 – 3.15. Miller units were plotted against time in one set of graphs and total expression values were plotted vs. time alongside the dilution corrected optical density of the culture at 600nm (open symbols) in another.

In the graphs where Miller units are plotted linearly against time genes/ORFs show the following patterns of expression. ORFs *ygjM* and *yncE* show higher levels of expression in the exponential phase of the growth curve, which slows as the cells enter the stationary phase of growth. The ORF *yhcN* shows low expression during exponential growth and increased expression during entry into stationary phase. The ORF *ycfR* is induced in early exponential phase followed by a drop in mid-exponential and induction again on entry into stationary phase. The ORFs *yceP, ydgH, yccV, yjdA, yegR, yedJ* and *ybhC* show higher expression levels in the exponential phases of growth while ORFs *hdeB, yahO, yhiM, hdeA, ygiN, yjfY, yraQ, yfjP, yjdI-K, yahLM, yigE, ybiM, yihR, yqhG, ybiJ, yeiN, ydhV, yjiW, ygaQ, ycjT, htrC, ypjC, ydeK, yegI* and *yfbL* show high expression levels during entry into the stationary phase of growth.

One of the attributes of plotting Miller units linearly versus time is that at high optical densities such as are found during entry into stationary phase, relative expression of a majority of genes appears to increase. This may be a consequence of expression of these genes continuing at the same rate while that of the bulk protein (i.e. ribosomes) decreases. To obtain a clearer picture of gene expression, total expression values were plotted on semi-logarithmic graphs versus time. Many ORFs show an increase in expression during growth over the whole period of the assay and are probably constitutively expressed; namely, *yceP, yhcN, yjdA, yhiM, ygiN, yahLM, yigE, yedJ, ybiJ, yegR, yihR, yccV, ydgH, yraQ, ybhC* and *yjiW*. Two ORFs namely *ygjMN* and *yncE* show a decrease in rate of expression as the growth rate decreases.

Strain K-12 MG1655 shows a reduction in growth rate after it reaches on optical density of between 0.3 and 0.4. This normally occurs at the time point of 150 minutes in figures 3.4 - 3.15. Many ORFs showed a marked increase in expression at this time point namely; *yahO, yfjP, yraQ, yqhG, ybiM, hdeB, hdeA, yhiM, yjfY, yihR, yjdI-K, ybiM, ydhV, yeiN* and *yfbL*. Of these, 3 ORFs (*yahO,* Ibanez-Ruiz, 2000*; hdeA* and *hdeB,* Bhagwat & Bhagwat, 2004) are known to be under the control of the stationary phase sigma factor (S) and it is possible that the other ORFs which show increased expression in the stationary phase of growth may be under the control of sigma S as well. This would have been tested had there been enough time.

The *ycfR* open reading frame was the only one which showed an increase in expression during the exponential phase of growth which slowed during the entry into stationary phase but again showed increased expression in late stationary phase. The ORFs *ycjT, htrC, ypjC, ydeK* and *yegI* were not plotted as lines on graphs showing total expression versus time as their expression values in the exponential phases were so low they could not be effectively plotted on logarithmic graphs. However their expression values in the stationary phase were high enough to be graphed and are shown as points on figures 3.14 and 3.15.

3.5.2.2. Growth tests on agar plates.

It was clear that the deletion of the genes selected did not result in slowed growth in LB broth, therefore, the growth tests listed below were carried out to detect any conditionally essential functions performed by these genes. The mutant strains were tested for growth on agar plates of different composition to detect any differences in rate of growth or number of colony forming units compared to the parent strain. *E. coli* has a variety of distinct and/or shared genetic pathways that enable it to survive the environmental stresses employed in these tests. The tests were designed to test the responses of all mutant strains and the parent to growth in response to high and low temperature, high and low pH, high osmolarity, anaerobicity, nutritionally starved media, metals and dye.

The growth tests carried out were: growth on LB agar at 30, 37 and 45 °C, growth on M9 glucose agar minimal medium at 30 and 37 °C, growth on LB agar in anaerobic chambers. Apart from these, mutant strains were tested for growth in the presence of varying levels of stress caused by heavy metals, osmolarity and dye. Two levels of stress were chosen (permissive and inhibiting) as any differences in fitness of mutant strains would be expected to show up compared to the parent strain for which the tests were standardised. These tests were; sensitivity to permissive and inhibiting levels of acidic (pH 5.8 & 5.6) and basic (pH 9 & 9.2) conditions, sensitivity to permissive and inhibiting concentrations of metal ions Co^{2+} ($CoCl_2$ 1 &1.5mM), Cu^{2+} ($CuSO_4$ 5 & 6 mM), Ni^{2+} ($NiCl_2.6H_2O$ 2 & 3 mM) and Zn^{2+} ($ZnCl_2$ 2 & 3 mM), salt (NaCl 0.8 & 1.2 mM) and crystal violet dye (10 & 20 μg/ml). Overnight cultures of all strains in LB broth were serially diluted in bacterial buffer in 96 well plates and 10 μl of each dilution was spotted onto the surface of agar plates using a multichannel pipette (Titertek).

All agar plates were incubated in appropriate conditions and were observed every 24 hours. Differences in spot or colony formation in terms of size and shape were noted and tests were stopped when the most dilute spot showed fully grown colonies. Colonies in

the two most dilute spots were then counted and averaged with both spots being given equal importance. Because of equal importance being given to the most dilute spot there is a large error margin in counts which is why rate of colony formation and number of colonies formed are taken into consideration in comparing growth of mutants to the parent. Experiments in which mutants showed different rates of growth or colony formation compared to the wild-type were repeated at least twice.

Incubation times varied between different conditions, plates of LB at 37 °C, M9 pH9, Cu^{2+} 5 mM, NaCl 0.8 mM, crystal violet 10 µg/ml were incubated for 24 hours. Plates of LB 30 °C, LB 45 °C, LB anaerobic, M9 37 °C, M9 30 °C, pH 5.8, pH 5.6, pH 9, pH 9.2, Co^{2+} 1 mM Co^{2+} 1.5 mM Cu^{2+} 5 mM Cu^{2+} 6 mM Ni^{2+} 2 mM Ni^{2+} 3 mM and Zn^{2+} 2 mM were incubated for 48 hours. Plates of M9 30 °C, pH 5.6 Ni^{2+} 3 mM, NaCl 1.2 mM and crystal violet 20ug/ml were incubated for 72 hours. Plates of Zn^{2+} 3mM were incubated for 96 hours and plates of Co^{2+} 1.5mM were incubated for 120 hours.

All colonies were counted after incubation and the number of colony forming units were compared to the parent MG1655 strain. For the majority of mutants there was no difference in growth compared to the parent strain in any of the test conditions. Colony numbers of all tests are provided in Appendix I. The colony numbers of MG1655 were counted and normalised to the number 1. All colony forming numbers for all mutants at the same condition were then averaged, compared to the normalised values for MG1655. Shown below in table 3.6 are average values of colony forming units of all mutants compared to the MG1655 parent strain and the corresponding standard deviation.

Table 3.6. Agar based phenotypic tests of mutants.

Condition	**MG1655 CFU**	**Average mutant CFU**	**Normalised CFU mutants**	**Standard deviation**
LB 37	$9x10^8$	$7.2 x10^8$	0.808	0.136
LB 30	$15 x10^8$	$7.7 x10^8$	0.519	0.340
LB 45	$12 x10^8$	$8.6 x10^8$	0.721	0.197

LB an	9×10^8	8.1×10^8	0.902	0.069
M9 37	10×10^8	5.5×10^8	0.559	0.312
M9 30	9×10^8	5.8×10^8	0.655	0.244
pH5.8	6×10^8	6.3×10^8	1.059	0.041
pH5.6	9×10^8	7×10^8	0.778	0.157
pH9	15×10^8	7.7×10^8	0.519	0.340
pH9.2	1×10^4	1.3×10^4	1.362	0.256
Co 1mM	7×10^8	4.9×10^8	0.710	0.205
Co 1.5mM	2×10^7	2.6×10^7*	1.319	0.226
Cu 5mM	16×10^8	11×10^8	0.694	0.216
Cu 6mM	5×10^8	2.9×10^8	0.596	0.286
Ni 2mM	6×10^8	5.2×10^8	0.883	0.083
Ni 3mM	6×10^8	3.8×10^8*	0.644	0.252
Zn 2mM	8×10^7	1.8×10^7	2.294	0.915
Zn 3mM	1×10^7	1.7×10^7	1.768	0.543
NaCl 1.2mM	8×10^8	2.8×10^8	0.357	0.455
NaCl .8mM	7×10^8	7.6×10^8	1.093	0.066
Crystal violet 10ug/ml	3×10^8	2.3×10^8	0.775	0.159
Crystal violet 20ug/ml	3×10^7	2.2×10^7§	0.735	0.187

*- *figures represent CFU's for all mutants other than* ΔyigE.

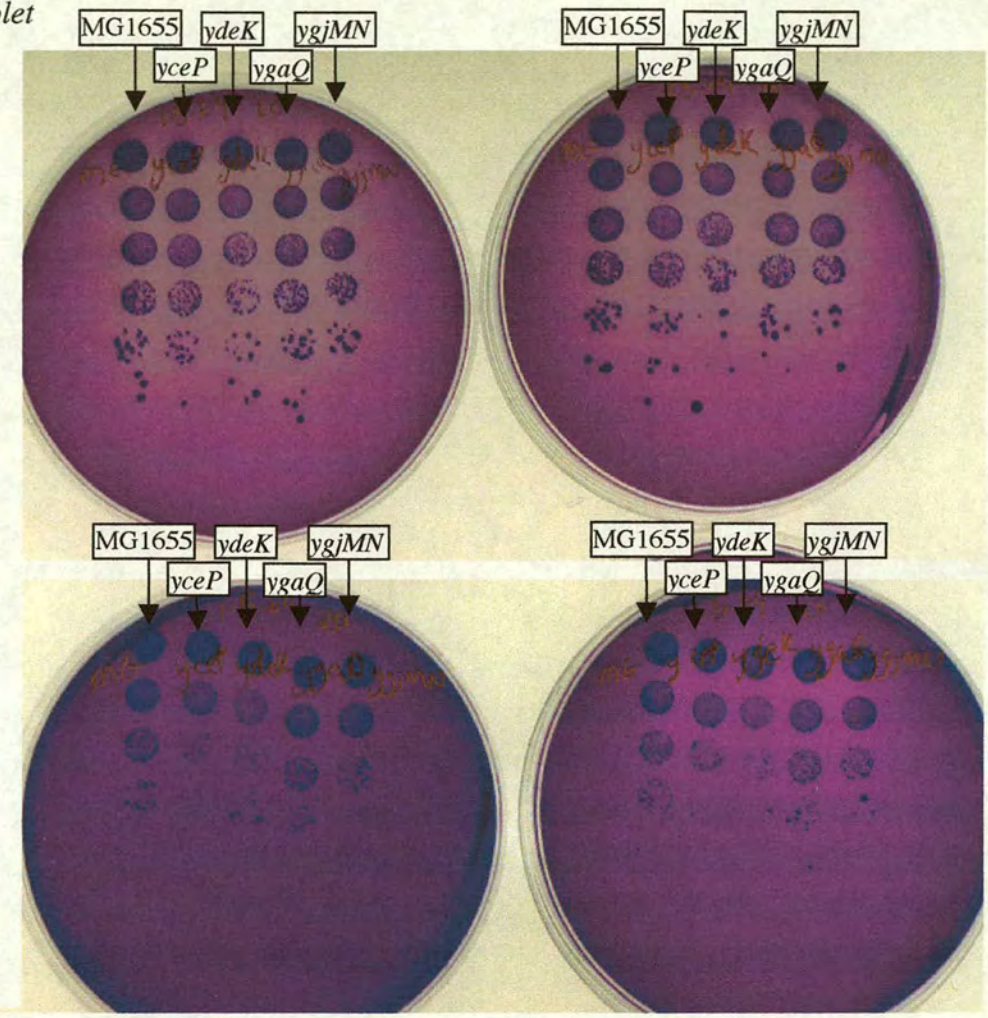
§- *figures represent CFU's for all mutants other than* ΔyceP, ΔydeK *and* ΔygjMN.

Any test where a mutant showed differences in growth compared to the parent strain was repeated at least twice to prove that the phenotypic difference was reproducible. The following mutants exhibited reproducible phenotypes.

The *yigE* mutant grew slower on LB plates containing restricting amounts of Ni^{2+} (3 mM) and Co^{2+} (1.5 mM). Nickel and cobalt affected only the growth rates and not the CFU of the *yigE* mutant as given enough time the CFU was comparable to that of the parent strain. This mutant was tested for growth in liquid medium containing nickel and cobalt and the growth rate was again slower than the parent strain. Other experiments showing the effects of complementing the *yigE* open reading frame in the mutant and parent strains, expression of *yigE* in the presence of nickel and cobalt and the effect of deleting *yigE* on the expression of the Mg^{2+}, Ni^{2+} and Co^{2+} transporter *corA* adjacent to it

have been carried out. The results are detailed in Chapter 6 dedicated to the functional analysis of the *yigE* open reading frame.

Figure 3.16. Colony formation on permissive (10 µg/ml) and limiting(20 µg/ml) amounts of crystal violet



Strains from left to right: MG1655 ΔlacZ, ΔyceP, ΔydeK, ΔygaQ *and* ΔygjMN.
Plates: Top row, 10 µg/ml crystal violet. Bottom row, 20 µg/ml crystal violet.

Figure 3.16 shows photographs of crystal violet (10 and 20 µg/ml) plates with strains MG1655 and mutants of ORFs *yceP*, *ydeK*, *ygaQ* and *ygjMN* after 72 hours of incubation at 37 °C. The plates show serial 1:10 dilutions of all strains spotted onto the surface of the agar medium with the least dilute at the top and most dilute at the bottom.

The tests were carried out in duplicate to show reproducibility between tests. Deletion mutants of ORFs *yceP*, *ydeK* and *ygjMN* show lower CFUs compared to the parent MG1655 strain on plates with inhibiting levels (20 μg/ml) of crystal violet dye. The *ydeK* mutant may be slightly more sensitive to the lower concentration (10 μg/ml) of crystal violet.

Growth of the *htrC* mutant was observed to be comparable to its parent at 45 °C. This came as an unexpected finding as mutations in the gene have reportedly resulted in severe temperature sensitivity as reported by Raina and Georgopolous (1990). The *htrC* gene is reportedly a heat shock gene whose deletion results in multiple phenotypes such as extreme temperature sensitivity, cell filamentation and constitutive overexpression of heat shock proteins (Raina & Georgopolous 1990). However the deletion mutant of the *htrC* gene in this study showed comparable viabilities at 45 °C to those of the parent strain. The *htrC* mutant was tested further to resolve the observed discrepancy and the results are reported in chapter 4.

3.6. Discussion:

This study was undertaken to identify and functionally analyse genes that are specific to *E. coli*. The precomputed BLAST analysis at the MBGD database enabled the identification of a set of 133 clusters that appear specific to the four *E. coli* genomes – K12 MG1655, 0157:H7, 0157:EDL933 and CFT073. To ascertain how the number of *E. coli* specific clusters has changed over the years since the first *E. coli* genome was sequenced the MBGD database was used to find clusters of genes specific to *E. coli* after the addition of each sequenced member of the gamma proteobacterial family in order of their sequence publication. This showed that the number of *E. coli*-specific clusters has been falling and where it initially stood at 2464 clusters (September 1997) it now stands at 133 when compared to 33 gamma-proteobacterial genomes. The fall in the number of species specific sequences is expected with an increase in the number of closely related members of the gamma proteobactialfamily being sequenced as this increases the

chances of finding homologs of specific genes. Sharp drops in the number of ORFan sequences in a genome when it is compared to a newly sequenced, closely related genome have been reported in a study comparing 60 published genomes and analysing dynamics of ORFan genes within these genomes (Siew & Fischer, 2003).

More than half (75) of the 133 clusters specific to *E. coli* have another *E. coli* specific cluster as an immediate neighbour suggesting a common event of gene acquisition or maintenance. While the majority of *E. coli*-specific clusters have no function attributed to them some, 23 out of 133 have known functions. The functions range from genes encoding enzymes (*aldA, sbm, agaB, agaD* and *gadA*), phage related genes (*dicC, dicB, fimC, fimE* and *fimH*), genes encoding membrane associated proteins (*uidC, uidB uidA* and *ompG*) and genes encoding regulatory functions (*tdcR* and *tnaC*). Some functions such as the resolvase activity of *rus* are well understood but other reported functions such as the regulation of viability in stationary phase encoded by *ssnA* remain poorly understood (Yamada et al, 1999).

One of the most important factors affecting the number of clusters that appear to be specific to *E. coli* in this study was in deciding which genome or genomes represents *E. coli*. The number of *E. coli* specific clusters is the highest when *E. coli* is represented by the single K12 MG1655 genome. Using combinations of K12 only, K12+0157:H7+0157:EDL933, K12+CFT073, K-12+0157:H7+0157:EDL933+CFT073 and K-12+*Shigella*2a (301 & 2457T) as representative *E. coli* genomes successively reduced the number of *E. coli* specific genes from 263 to 239 to 153 to 133 and 117. This suggests that the K12 strain of MG1655 shares the majority of its ORFan genes with the two 0157 strains and shares far fewer ORFans with the CFT073 strain or the *Shigella* species.

All four strains of *E. coli* are very different sharing only 39.2% of their genes, with the pathogenic genomes being as different from each other as they each are from the non-pathogenic strain (Welch et al, 2002). Maybe occupying diverse niches has led to the

loss of some genes specific to *E. coli*. In the case of *Shigella* it might be due to the exceptionally high numbers of IS elements in this genus that may have led to the deletions or insertional inactivation of *E. coli* specific sequences (Jin et al, 2002).

Of the 49 genes deleted in this study 21 remain specific to *E. coli* while others show homologues in closely related gamma proteobacterial genomes or weak homologues in non-gamma proteobacterial genomes that were published after these genes were selected for deletion. Although none of the genes deleted caused slowed growth in LB broth these genes are expressed and show patterns of expression which vary in different phases of growth. Of the 39 gene expression assays carried out, 14 ORFs showed relatively higher levels of expression in the exponential phase of growth compared to their expression during early stationary (defined as the transition phase between exponential and stationary phases) or stationary. 28 of 39 ORFs were upregulated during the early stationary phase and 9 of 39 genes were upregulated in the stationary phase. The substantially higher number of ORFs showing increased expression during entry to stationary phase may be because their expression is under the control of the stationary phase transcription factor sigma S. Although this needs to be experimentally verified it would be of interest to find how many *E. coli* specific ORFs are truly under the control of sigma S.

Growth and stress tests on agar based media showed that most mutants have growth rates and CFUs comparable to that of the parent strain. This may be due to the relatively small (22) range of tests that were carried out. A system of methodical phenotypic tests such as those described in the Biolog (Bochner, 2003) system would possibly result in more phenotypes being discovered. However even with the narrow range of tests employed here there were mutants that showed altered growth rates and/or colony forming unit numbers.

Deletion of the *yigE* ORF resulted in a mutant showing increased sensitivity to inhibiting levels of nickel and cobalt (3 and 1.5 mM respectively) compared to the parent MG1655

strain. Further tests on the *yigE* mutant are detailed in chapter 6. Growth of mutants in the presence of the dye crystal violet uncovered three mutants with decreased CFUs compared to the parent strain. Crystal violet is a basic dye that is normally excluded from the cell in K12 strains with functional cell envelopes. In mutants that have defective cell envelopes such as mutants of gene *lpxC* (formerly *envA*, involved in biosynthesis of lipid A) crystal violet enters the cell and disrupts protein synthesis by binding to ribosomes (Gustaffson et al, 1973).

One of three mutants sensitive to crystal violet lacks the *yceP* ORF. This ORF lies in the *solA-dinI* intergenic region and has no functions associated with it. It has been reported to be up-regulated during heat shock (Richmond et al, 1999), oxidative stress (Zheng et al, 2001), and acclimatization at low temperature (Polissi et al, 2003) by different groups. It may be that *yceP* functions as a general stress response protein. Although it has been reported to be induced during heat shock the *yceP* mutant showed growth and CFUs similar to the parent during growth on LB agar at 45 °C. The encoded protein has no predicted transmembrane domains which suggests that its function is carried out in the cytosol. It may perhaps protect ribosomes and other structures from crystal violet or other basic dyes.

The second mutant that shows sensitivity to crystal violet is the *ydeK* deletant. This ORF codes for a potential 1325 amino acid protein that has a predicted lipid anchor suggesting its subcellular location to be associated with the membrane. This ORF bears weak homology to fungal mitochondrial import site proteins ISP42 and MOM38, which are essential for the import of protein precursors into mitochondria. Perhaps *ydeK* forms part of an *E. coli* import site/complex that functions to maintain a lipid barrier to dyes such as crystal violet.

The last mutant which shows sensitivity to crystal violet has the ORFs *ygjM* and *ygjN* deleted. The ORFs code for predicted proteins of 138 and 104 amino acids respectively. The *ygjN* ORF has no predicted motifs, while the *ygjM* ORF has one HTH cro/C1-type

DNA-binding domain. This binding domain is conserved in several other transcriptional repressors such as the purine repressor (PurR), the lactose repressor (LacI) and the fructose repressor (FruR). Deletion of this putative repressor may perhaps have increased the import of crystal violet into the mutant cell.

Perhaps the first question that needs to be asked is the intracellular level of crystal violet in each of the above mutants and how this compares to the parent strain. Any mutation that disturbs the lipid barrier of *E. coli* to a dye such as crystal violet may have potentially important therapeutic applications. These genes perhaps encode mechanisms that *E. coli* uses to protect itself against therapeutic agents such as crystal violet.

The actual number of genes specific to *E. coli* may actually be quite different to that reported here since the *E. coli* is a very diverse species whose genomes can vary by up to 1 megabase (Bergthorsson and Ochman, 1998). This study however sheds some much needed light on the growing family of genes that are species specific. It highlights the dynamic nature of its member genes in that with every new genome sequenced homologs of previously species specific genes may be found. However there remain a substantial number of genes that are still specific to *E. coli* and, as shown here, are transcriptionally active. The transcription patterns vary with the growth phase and deletion of some of these genes produces clear phenotypes which point to their possible physiological activity. One of the most interesting avenues of research is in understanding the role played by species specific genes in pathogenesis since this may lead to novel therapeutic targets and help reduce the use and associated dangers of broad spectrum antibiotics.

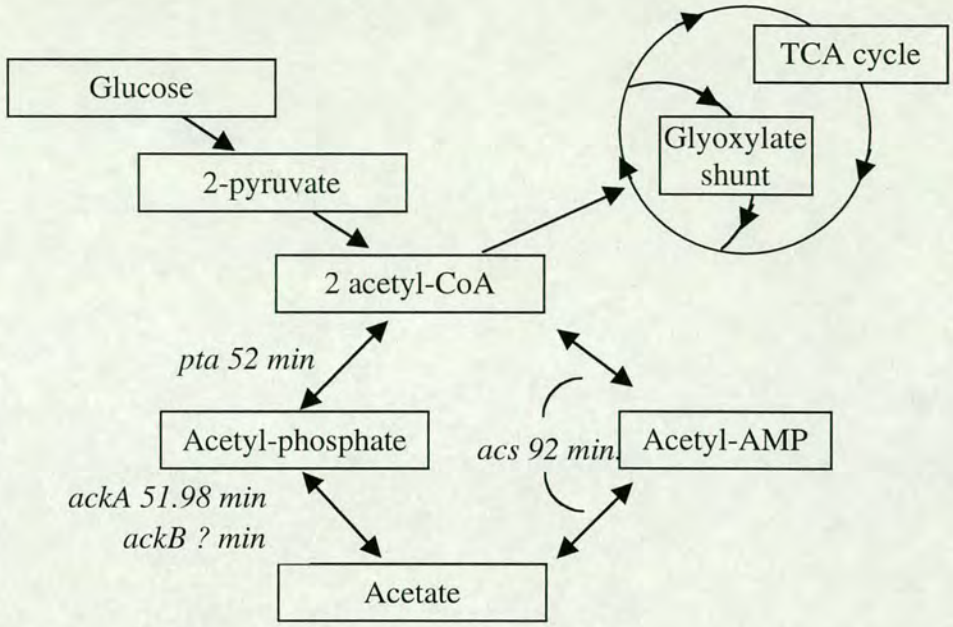
Chapter 4: Chromosomal location of the *ackB* gene.

4.1: Introduction:

The approach to functionally analysing genes on the MG1655 chromosome used so far in this study has been to choose predicted open reading frames that are specific to *E. coli*, delete them and test mutants for consequent effects. Close to 2.1% of named mutant phenotypes on *E. coli* genome do not have ORFs or genes associated with them (Serres et al, 2001). Linking these reported phenotypes with associated genes is essential in reconciling past *E. coli* genetic research with our current understanding of the *E. coli* chromosome. This chapter details linking one such reported phenotype to its gene on the chromosome. The phenotype selected was the inability of a mutation designated as *ackB* to grow on a medium containing acetate as the sole carbon source.

E. coli can use acetate as the sole carbon source for growth, amino acid production and lipid synthesis. To use acetate as a carbon source it first needs to be converted to acetyl-CoA which is then taken up into the glyoxylate shunt or Krebs cycle. There exist two known pathways that convert acetate to acetate-CoA. One pathway is encoded by the *ackA-pta* genes found at the 50 minute region of the *E. coli* genome. The protein acetyl kinase (AckA) first converts acetate to acetyl-phosphate with the loss of one ATP. Acetyl phosphate is then converted to acetyl-coA by phosphotransacetylase (Pta) (Kakuda et al, 1994). The acetyl-CoA synthase (acs) gene located at 92 minute on the *E. coli* map encodes the second pathway. Acs first converts acetate to acetyl-AMP and then converts acetyl-AMP to acetyl-CoA. The acetyl-CoA is then shunted into the TCA cycle or the Krebs cycle as demonstrated in figure 4.1.1. (Brown et al, 1977).

Figure 4.1. Acetate and the glucose fermentation pathway.



The two pathways for activating acetate exist as high and low affinity systems. The *acs* pathway is a high affinity activator which functions when concentrations of acetate in the environment are low. The *ackA-pta* system is a low affinity activator of acetate which functions when acetate concentrations in the environment are high. Deleting all three genes together results in a mutant unable to grow at low or high concentrations of acetate as the sole carbon source (Kumari et al, 1995).

E. coli has another reported acetate kinase gene named *ackB* whose exact genetic position remains unknown. According to published literature the *ackB* gene product performs a function similar to that of the *ackA* gene located at 50 minutes on the K12 chromosomal map. Both gene products reportedly carry out the enzymatic conversion of acetate to acetyl phosphate. Although the position of *ackA* and the enzymatic characteristics of AckA are well studied (LeVine at al, 1980, Kakuda et al, 1994) those of *ackB* and its encoded product remain poorly understood. *AckA* is a gene well conserved among bacterial genomes; it is curious that *E. coli* appears to have a second acetate kinase that has not been reported in any other genome and which is required in

addition to AckA. This study was carried out to identify the gene whose disruption would cause the reported *ackB* phenotype described by Pascal et al (1981). If *ackB* is a gene specific to *E. coli* and dissimilar to *ackA* it would be of interest to know why a mutation in *ackB* affects the acetate kinase activity of a functioning *ackA* gene.

The *ackB* mutant was identified amongst others in a study where mutants were generated using N-methyl-N-nitroso N-nitrosoguanidine (MNNG) from a strain carrying a mutation in the *adhE* (*ana*$^-$) gene with the aim of isolating mutants that were unable to use exogenous electron acceptors such as nitrite or nitrate for growth in anaerobic glucose minimal medium. The *ackB* mutant strain LCB190 was unable to grow on minimal medium with acetate as the sole carbon source. The strain was also negative for H_2 gas production from pyruvate and showed 12-33% lower nitrite reductase activity when glucose was the electron donor (Abou Joude et al 1978).

In a separate report LCB190 was shown to lack acetate kinase activity, to grow slower in glucose medium under aerobic conditions and not to accumulate acetate compared to the *ackB*$^+$ parent. The mutation was mapped using conjugation to the 39 minute region on the *E. coli* map (Pascal et al 1981) and was pronounced distinct from the acetate$^-$ mutant (*ackA)* mapped at 50 minutes by Brown et al (1977). These studies have led to the currently accepted listing of acetate kinase genes *ackA* and *ackB* in *E. coli.*

4.2. Transduction of ackB mutant to *acetate*$^+$.

The first step to locate the *ackB* mutation was to verify if the mutant LCB190 could be transduced to an acetate$^+$ phenotype using P1 lysates prepared from wildtype MG1655. Strain LCB190 was transduced using a P1 lysate prepared from MG1655 in late logarithmic phase as detailed in Materials and Methods (2.18). Transduction of leucine (Leu$^+$) was used as a positive control for the transduction. 250 µl of the transduction mix were plated on M9 minimal medium containing acetate (0.2% sodium

acetate) as sole carbon source and separately M9 glucose minimal medium lacking leucine. All plates were supplemented with thiamine, threonine, 1 M $MgSO_4$ and 0.5M $CaCl_2$ and incubated for 4-5 days at 37 °C. Table 4.1. shows the results from the transduction reactions.

Table 4.1. Number of acetate+ and leucine+ LCB190 transductants.

Phenotype	MG1655 lysate	Phage free control
Acetate$^+$	20	0
Leucine$^+$	400	0

The transduction frequency to acetate$^+$ was 20 times lower than of leucine$^+$. The reasons for such low transduction frequency may be due to either the chromosomal position or the nature of the *ackB* mutation. Transduction frequencies using P1 lysates can vary for markers depending on their chromosomal position (Masters, 1977) or the recombination efficiency of the host (Newman and Masters, 1980). Low transduction frequencies to acetate$^+$ may also be due to the nature of the *ackB* mutation which was made using the mutagen MNNG. This mutagen creates random base substitution mutations in a dose dependent manner near the replication fork (Myung & Kolodner, 2003) and it is possible that the AckB phenotype is caused by two or more relatively close yet unique mutations which have to be corrected together.

4.3. Are ackA *(ackA202) and* ackB *two distinct mutations?*

The next test was carried out to determine if the *ackB* mutation was distinct from the *ackA* mutation mapping at 50 minutes. To determine this P1 lysates were prepared from the two acetate$^-$ strains obtained from Dr. Mary Berlyn (CGSC database). The resulting lysates from strains LCB190 and ackA202 were used to transduce the two acetate$^-$ mutants ackA202 and LCB190 respectively. This experiment was carried out to resolve

if the acetate$^-$ phenotype of the two strains was due to mutations in the same gene, and if so, no acetate$^+$ transductants would be recovered. 250 µl of the transduction mix were plated on M9 acetate plates as above and incubated for 4 days at 37 °C. and the colony numbers are shown in table 4.2. below.

Table 4.2. Number of acetate+ ackA202 and LCB190 transductants.

Recipients	AckA202 lysate	Lysate free control	LCB190 lysate	Lysate free control
LCB190	10	0	0	0
AckA202	0	0	17	0

Positive transduction controls threonine$^+$ for LCB190 recipient and histidine$^+$ for AckA202 recipient showed between 90-100 colonies. The transduction frequencies of acetate$^+$ were again seen to be very low for both strains, however the lack of any colonies on the lysate free control plates discounts the possibility of the putative transductants being revertants rather than true transductants. The very low transductant numbers obtained from the above experiment make it difficult to decide whether the two mutations are distinct or identical. The next set of transductions were carried out to test if the *ackB* mutation maps close to the 39 minute region as claimed by Pascal et al (1981).

4.4. Verifying the claimed 39 minute position of the ackB *mutation.*

To verify the reported 39 minute chromosomal position of *ackB,* lysates of *E. coli* strains carrying Tn*10* transposons at positions 38.3, 39.5, and 40.3 (Nichols et al, 1998) were used, to test the linkage between the transposon encoded tetracycline resistance (*tet*^{R}) and acetate$^+$ phenotype. Transduction of leucine$^+$ and *tet*^{R} were used as positive controls and transduction numbers are shown in table 4.3.

Table 4.3. Numbers of acetate$^+$, leucine$^+$ and tetracycline resistant LCB190 transductants.

Lysate	Acetate$^+$	Leucine$^+$	tet^R	Acetate$^+$, tet^R
CAG12151 (38.3)	45	233	475	0
CAG18464 (39.5)	42	243	271	0
CAG18465 (40.3)	41	237	386	0

The transduction reactions showed that although each CAG strain was capable of transducing LCB190 to acetate$^+$ and that the acetate transduction frequencies were as low as those found with MG1655 lysates, there was no linkage between the tetracycline markers at positions 38.3, 39.5 and 40.3 and acetate$^-$ phenotype of the *ackB* mutation. The reason no acetate$^+$ and tetracycline resistant colonies were recovered may be due to two possibilities. One possibility may be that selecting for these two phenotypes together might be physiologically or genetically difficult due to some unknown reason. Since no tetracycline resistant transductants were tested for their ability to grow on M9 acetate this possibility remains untested. The other and more probable possibility for not obtaining *tet*^{R} acetate$^+$ transductants was that the *ackB* mutation is not linked to the tetracycline markers at map positions 38.3, 39.5 and 40.3. The experiments described next were carried out with the aim of mapping *ackB* on the chromosome of strain LCB190.

4.5. Mapping the ackB mutation on the chromosome of mutant LCB90.

A random transposition library was then employed to try to isolate a strain in which the *ackB* mutation would be linked to a selectable phenotype (Km^R) on the transposon. The transposon used was the mini-Tn*10* derivative 103 marked with a kanamycin resistance (kan^R) fragment from Tn*903*. The two ends of the transposon have perfect inverted repeats of 70 bp carrying the outside end of IS*10* right. The delivery vehicle is a P_{am} 80 lambda hop phage which carries a *Ptac-ats1 ats2* transposase gene in cis (Kleckner N., 1991). The advantage of using this system is that since the transposase is carried on the phage it ensures stable insertions on the chromosome. A schematic of the transposon mutagenesis and subsequent experiments to map *ackB* on the chromosome of strain LCB190 is shown in figure 4.2.1.

The first step was to create a library of random transposon inserts on the K-12 MG1655 chromosome. A late log phase culture of MG1655 was infected with the lambda phage vehicle and incubated overnight at 37 °C. The overnight culture was then used to prepare P1 lysate. The resulting P1 lysates were then used to transduce kan^R into strain LCB190 followed by screening kanamycin resistant clones for the $acetate^+$ phenotype. All clones which were kan^R and $acetate^+$ were then checked for $leucine^-$ and streptomycin sensitive phenotypes of the parent LCB190 strain.

There were 21 $acetate^+$ kan^R LCB190 clones obtained from two separate rounds of P1 transductions from the mini-Tn*10* library. Fresh P1 lysates were prepared from all 21 $acetate^+$ clones and the resulting lysates were used to transduce the parent LCB190 ($acetate^-$ kan^S) clones to kan^R and a hundred kan^R clones were patched on M9 glucose and M9 acetate agar media. Transduction frequencies of $acetate^+$ transduction were calculated as a percentage of Kan^R clones and are shown in table 4.2.1.

Figure 4.2. Mapping the ackB *mutation in strain LCB190*

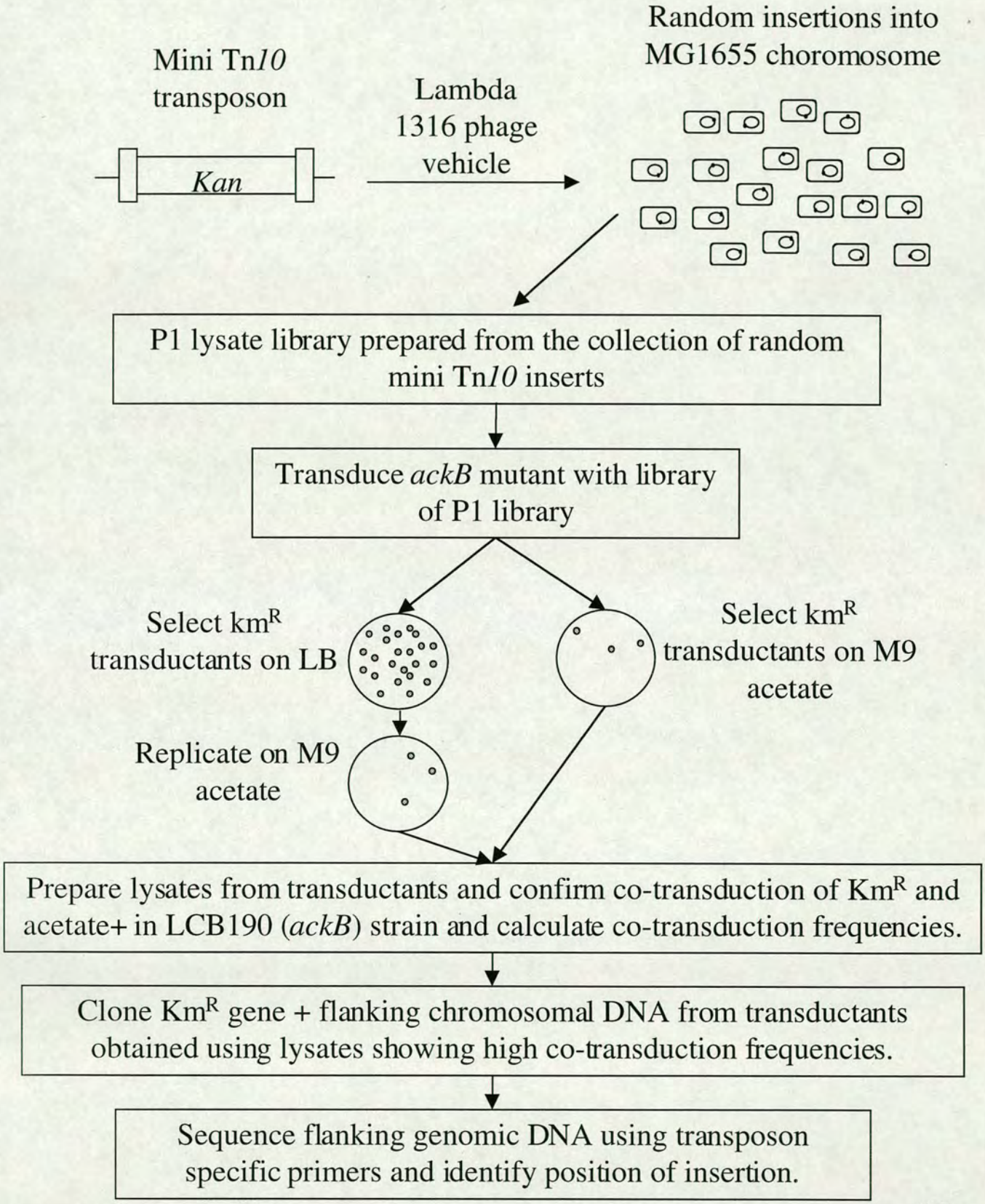


Table 4.4. Co-transduction frequencies of kanamycin resistance and acetate$^+$ of 21 mini-Tn10 mutants.

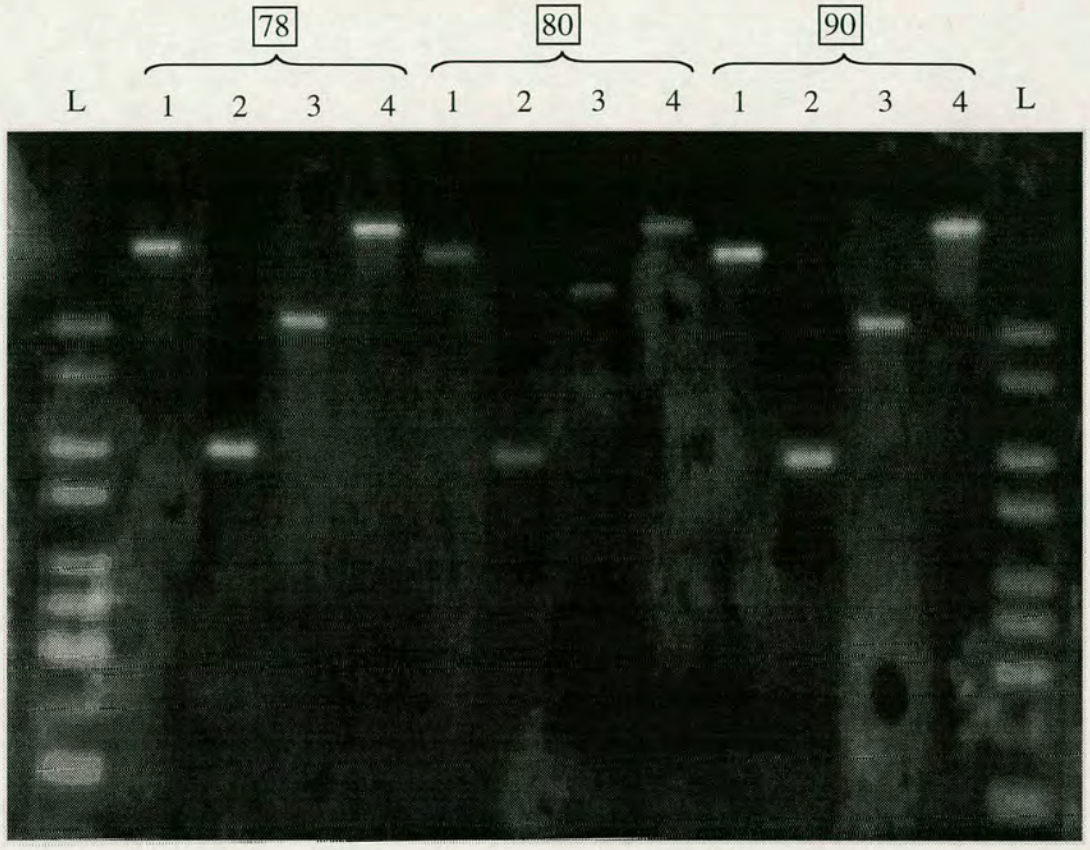
Strain number	% Co-transduction frequency
76	23
77	11
78	86
79	10
80	82
81	3
82	10
83	10
84	10
85	27
86	16
87	8
88	20
89	33
90	86
91	10
92	-
93	22
94	20
95	35

Three clones namely 78, 80 and 90 which showed co-transduction frequencies between acetate$^+$ and kanamycin resistance of 86, 82 and 86 percent respectively were selected for further genetic analysis.

To map the position of the mini-Tn*10* transposon on the chromosomes of the three selected strains, the transposon along with the flanking DNA had to be first cloned and then sequenced. To ensure that the strains all had single transposon inserts and that the restriction fragment would contain enough flanking chromosomal DNA to identify the insertion point, chromosomal DNA fragments resulting from restriction digestion using various enzymes were fractionated by agarose gel electrophoresis and used in a Southern blot where the kanamycin resistance gene was used as a probe.

Chromosomal DNA of all 3 strains was prepared and 5 µg of DNA from each mutant was digested individually with 20 units of enzymes *Eco*R1, *Eco*RV, *Pst*I and *Sal*I supplied by NEB. The digested DNA was fractionated on a 0.7% agarose gel overnight and the fragments were blotted on to a nitrocellulose sheet and the fragments were probed with radiolabelled *kan*^{R} gene of the mini-Tn*10* as described in Materials and Methods (section 2.11). Shown in figure 4.2.2. below is the image obtained after scanning the phosphor imager screen exposed to the probed nitrocellulose sheet. The figure shows that the clones 78 and 90 have the mini-Tn*10* transposon in identical positions on the chromosome and both are likely to be the progeny of a single insertion event. Clone 80 shows very similar banding for all lanes except for the *Pst*I lane where the hybridising fragment appears to be slightly larger compared to those of clones 78 and 90.

Figure 4.3. Southern hybridization of Eco*RI,* Eco*RV,* Pst*I and* Sal*I fragments of clones 78, 80 and 90.*



L: 1 kb DNA ladder supplied by Fermentas™. Fragment sizes from top to bottom in kb: 10, 8, 6, 5, 4, 3.5, 3, 2.5 and 2.

1: Eco*RI digested DNA; 2:* Eco*RV digested DNA; 3:* Pst*I digested DNA; 4:* Sal*I digested DNA. Boxed numbers indicate source of chromosomal DNA.*

To sequence the smallest chromosomal fragments (*Eco*RV and *Pst*I) 5 µg of chromosomal DNA from clones 78 and 80 were redigested and ligated to plasmid pBR328. Overnight ligation reactions were used to transform competent DH5α strains of *E. coli* and transformants were selected on LB kanamycin plates. Kanamycin resistant clones emerging from ligations of *Eco*RV with pBR328 were checked for tetracycline

sensitivity and those emerging from ligations of *Pst*I fragments with pBR328 were checked for ampicillin sensitivity.

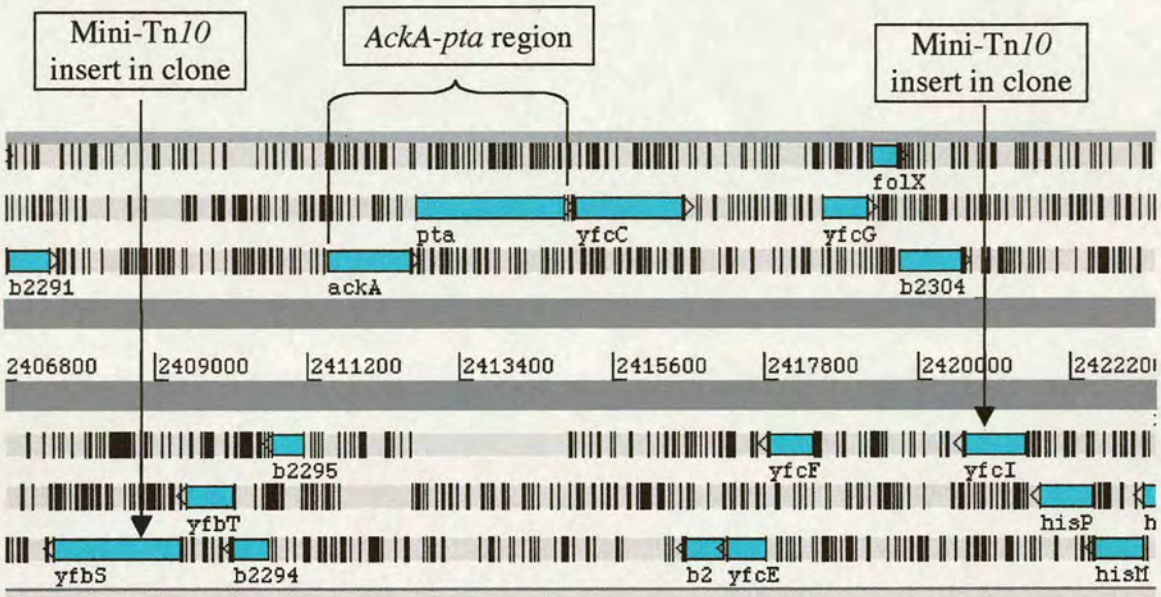
Four Km^R Amp^S *Pst*I:pBR328 transformants for strain 78 and three Km^R Tet^S *Eco*RV:pBR328 transformants for strain 80 were purified and plasmids prepared from these clones were sequenced as recommended by the ABI377 manual. The primers used in the sequencing reactions (Leftinsiderep and Rightinsiderep, Table 2.1 Materials and Methods) are specific for the mini-Tn*10* transposon and are designed just inside the *rep* region so that the resulting sequence would have the entire repeat region and the chromosomal sequence flanking the point of insertion.

Sequencing the four *Pst*I:pBR328 plasmids of transductant mutant strain 78 with primer Left Inside Rep showed 401 base pairs of the repeat region and immediately after 299 bp which were 99% identical to bp 2408755-2409054 of the MG1655 chromosome. Sequencing with the primer Right Inside Rep showed 476 bp of transposon sequence and 172 base pairs of sequence 99% identical to base pairs 2408764-2408592 of the K-12 chromosome. This showed that in the case of clone 78 the mini-*Tn10* transposon was integrated in ORF *yfbS*.

The three *Eco*RV:pBR328 plasmids of strain 80 sequenced with primer Left Inside Rep produced 401 bp of transposon sequence and 295 bp matching positions 2420889-2421184 of the K-12 chromosome. Sequencing the plasmids with the Right Inside Rep primer produced 476 bp of mini-*Tn10* transposon sequence and 153 bp of sequence which matched bp 2420896-2420743 of the K-12 chromosome. This places the transposon of strain 80 within the *yfcI* ORF.

In all seven sequenced clones, the site of insertion of the mini-Tn*10* transposon showed a consensus of CTGGG. The ORFs *yfbS* and *yfcI* lie at 51.9 and 52.18 minutes on the K-12 chromosome. The proximity of mini-Tn*10* inserts to the *ackA-pta* region is shown in figure 4.2.3. below.

Figure 4.4. Mini-Tn*10 inserts in mutants 78 and 80 and their proximity to* ackA-pta.



The two grey bars represent positive (above) and negative (below) strands of DNA. The six lines above and below them serve to represent the six possible coding frames. Known and predicted genes are shown in blue in their respective coding frames. Black vertical lines represent possible stop codons. Positions of mini-Tn*10 inserts in clones 78 and 80 are indicated by arrows.*

The close proximity of the *kan*^{R} linked acetate^{+} phenotype to the *ackA-pta* genes suggests a link between the acetate^{-} phenotype of strain LCB190 and the *ackA* gene at 51.98 minutes. Since the acetate^{-} phenotype is so closely linked to a gene known to encode an acetate kinase it is likely that the *ackA* gene in strain LCB190 is mutated. This study found no likely evidence for the existance of the *ackB* gene.

4.6. Discussion

The *ackB* mutation described by Pascal et al (1981) renders the mutant unable to grow in medium where acetate is the sole carbon source. Transduction experiments using phage P1 carried out here have shown that the *ackB* mutant LCB190 can be transduced to

acetate$^+$ using lysates of wild type MG1655. Transduction frequencies of acetate$^+$ are low compared to markers such as leucine$^+$ or threonine$^+$ and this may perhaps be due to the nature of the *ackB* mutation. Perhaps the *ackB* mutation is a result of more than one base substitutions in the target gene or the chromosomal position of the *ackB* mutation makes transductions less efficient.

While the nature of the *ackB* mutation remains unresolved this study has shown that acetate$^+$ phenotype can be transduced from the two acetate mutants *ackA202* (*ackA* mutation) and LCB190 (*ackB*) to each other at transduction frequencies that are again low. The low transduction frequencies make it difficult to judge by transduction experiments alone the uniqueness of the two mutations. Two different research groups created the two mutants using different methods. The *ackA* mutant *ackA202*, was isolated by selecting resistance to fluoroacetate (LeVine et al, 1980) while the *ackB* mutation was accomplished using the mutagen N-methyl-N-nitroso N-nitrosoguanidine (Pascal et al, 1981). The *ackA202* mutation is less stable on agar plates where mutants are liable to revert to acetate$^+$ phenotypes (LeVine et al, 1980). The *ackB* mutation on the other hand is very stable and shows no revertants. It may be possible that the *ackA* and *ackB* mutations affect a single gene or genetic region in different positions.

The initial experiments were carried out to confirm the claimed 39 minute map position of *ackB* as reported by Pascal et al (1981). Transduction experiments carried out in this study have shown no linkage between the acetate gene and the 38.3-40.3 region of the *E. coli* chromosome. This finding led me to employ a random transposon mutagenesis approach to identify the position of the *ackB* mutation by first using a random library of mini-Tn*10* inserts on the MG1655 chromosome, transducing the LCB190 strain to acetate$^+$ and then sequencing outward from the transposon.

Using the transposon mutagenesis approach the *ackB* acetate$^-$ phenotype has been mapped close to the 52 region of the *E. coli* chromosome. Being mapped so close to the *ackA-pta* region it is possible that the *ackB* gene may be the same as *ackA*. While the

possibility of a mutation in gene *ackA* in strain LCB190 has not been verified here, the activity of the AckA enzyme in this strain has been tested by Pascal et al. The strain shows very low acetate kinase activity compared to the parental strain LCB900, 0.01 μmol min^{-1} (mg protein)$^{-1}$ compared to 1.6 μmol min^{-1} (mg protein)$^{-1}$. The activity of enzyme Pta in strain LCB190 and the parent strain were reported to be identical at 1.3 μmol min^{-1} (mg protein)$^{-1}$. This seems to suggest that the strain LCB190 may have a defective *ackA* gene or have another defective gene close to *ackA* that is necessary for its activity.

Any experiments carried out in the future to resolve this would have to involve sequencing the entire *ackA-pta* region along with its flanking chromosomal region. This work was not carried out in this study due to time constraints. However these experiments would perhaps highlight the genetic regions which when mutated render the AckA enzyme inactive or uncover another gene/enzyme close to *ackA* on the chromosome whose activity is necessary for a functioning AckA enzyme.

Chapter 5: *HtrC*: Heat shock gene or a new ORFan?

5.1: Introduction:

The *htrC* gene was first described by Raina & Georgopoulos (1990). The authors reported the isolation of two Tn*5* transposon mutants that were hypersensitive to a normally sublethal heat shock at 50 °C. These mutants were described as showing extensive filamentation and lysis at temperatures above 42 °C. The gene responsible was reported to have been identified by complementation of the temperature sensitive phenotype by the *htrC* gene encoded by genomic DNA in cosmid, plasmid and lambda phage libraries. The *htrC* gene was mapped in Southern blot experiments using overlapping lambda clones of *E. coli* genomic DNA as targets and *htrC* complementing cosmids as probes. This was followed by cloning of the transposon containing gene and sequencing which showed that the transposons were inserted at positions 131 and 357 in a 537 bp open reading frame.

The authors constructed a null deletion of the now-identified *htrC* gene which displayed the same phenotype as the Tn*5* insertion mutants. P1 transduction of the null *htrC* mutation into different strains of *E. coli* showed that the phenotype was not strain specific. The 21 kDa *htrC* encoded protein was predicted to be very basic and the authors were unsuccessful in overproducing it possibly due to its high percentage of rare codon usage. The transcript of the *htrC* gene was mapped and reported to be about 550-600 nucleotides long and spanning the *htrC* open reading frame of 537 bp. The transcriptional start site was located 33 nucleotides upstream of the putative AUG initiation codon. Using a plasmid carrying the *htrC* ORF the authors also reported that the level of the RNA transcript of *htrC* increased upon a shift from 30 to 42 °C but was most abundant at 50 °C.

Following this it was further reported that the *htrC* transcript was dependent upon the heat shock sigma factors sigma 32 and sigma E for expression. Using two dimensional PAGE electrophoresis and ^{35}S labeled total protein extract from *htrC*$^{+}$ and *htrC*$^{-}$ cells the authors found that the *htrC* mutant constitutively overproduced heat shock proteins such as DnaK, GroEL, GrpE, HtpG, and also other non-heat shock genes. Between the temperatures of 39 and 41 °C the *htrC* mutant filamented extensively. The other phenotypes of the *htrC* mutation reported in the same paper were as follows. The *htrC* mutation when combined with a *lon* deletion reportedly abolished the UV sensitivity caused by *lon* but the double mutant retained the temperature sensitive phenotype of the single *htrC* mutant. *HtrC* mutants were also reported to be defective in cellular proteolysis and to degrade puromycyl peptides at a reduced rate.

The varied phenotypes reported led the authors to the following conclusions about the function of the HtrC protein. Since the mutant showed constitutive overexpression of heat shock genes it was proposed that HtrC directly or indirectly modulates the expression of sigma 32. The slower breakdown of puromycyl polypeptide led to the theory that HtrC might itself be a protease or may modulate the activity of other proteases. The protease theory was also used to explain why sigma 32 modulated genes were overexpressed, as sigma 32 has a very short half-life and in the absence of protease activity, genes under its control would be abnormally expressed. Since the *htrC* mutant also formed filaments at temperatures over 42 °C it was suggested that *htrC* plays a role in cell division and perhaps interacts with the *ftsZ* gene product. The protein was also reported to show homologies to the *B. subtilis* sigma factor SpoIIA and to the hsp26 family of eukaryotic low molecular weight heat shock proteins and some serine proteases.

Since the original publication describing the role of *htrC* in the heat shock response of *E. coli* there has been little further work on this gene and its function. The large number of genomes sequenced so far do not show any significant homologs of *htrC*. The gene appears to be specific to *E. coli* and a comparison of MG1655 and 0157:H7 shows the

C-terminal portion of *htrC* to be missing in the pathogenic strain. The only other gene with significant *htrC* homology outside the *E. coli* genomes is a hypothetical protein of 143 amino acids in *Pseudomonas syringae pv. maculicola* which shows 32% identity to residues 47-148 of HtrC.

HtrC was selected as a deletion target in this study because it was one of very few *E. coli* specific genes that had an assigned function. While the gene reportedly played no role in the growth of the cell, its function was conditionally essential during heat shock. Many questions about the role of *htrC* during heat shock remaine unanswered. Its role as a protease regulating the expression of heat shock genes also remains to be understood. Answering the broader question of why *E. coli* would have a species-specific regulator of heat shock genes is important in understanding the evolution of the heat shock response in *E. coli.*

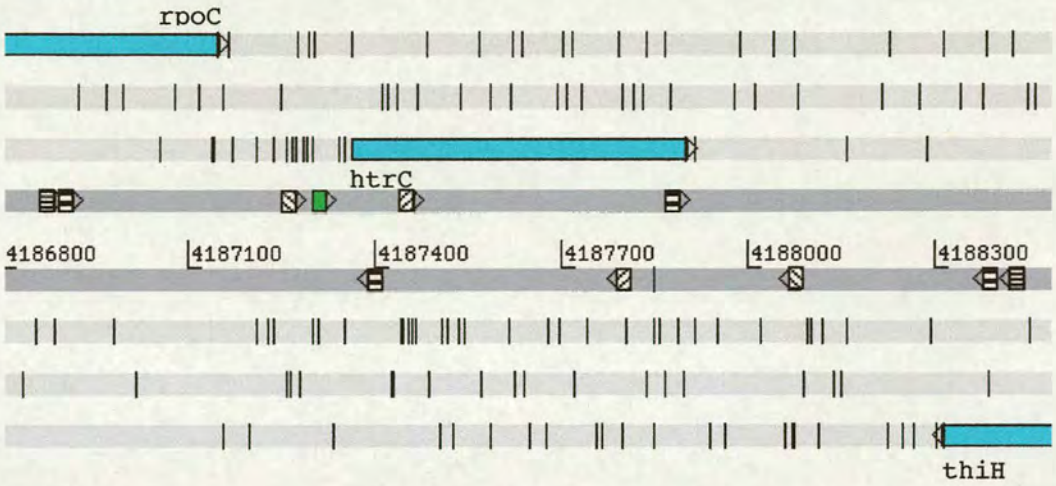
5.2: Results

5.2.1: Functional analysis of the *htrC* gene product:

The MG1655 *htrC* gene is in the *rpoC-thiCEFSGH* intergenic region at 4187.4 kb on the chromosome. It is a 540 bp gene with codon usage index 3 encoding a putative small 179 amino acid, very basic protein. Its predicted molecular mass and isoelectric point are 20987.2 kDa and 9.5 respectively.

Deletion of the *htrC* gene was carried out as described earlier. Figure 5.1 shows a schematic representation of the *htrC* region and the primer positions used for gene deletion and confirmation.

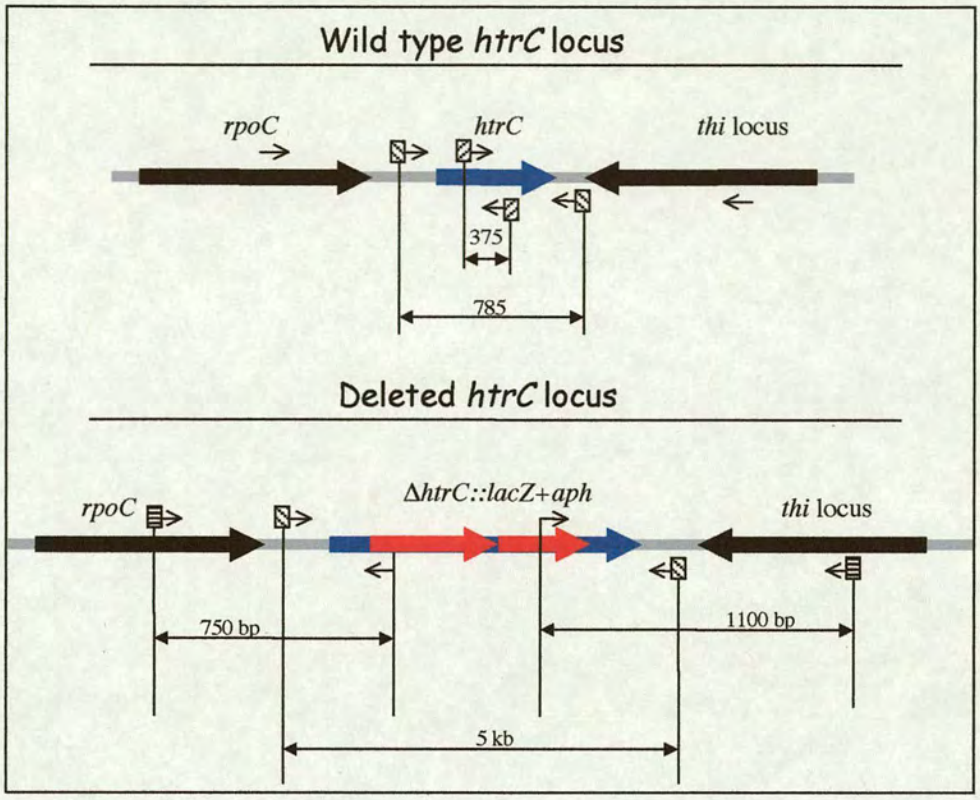
Figure 5.1. Primers used for htrC *deletion and confirmation.*



The two grey lines represent positive and negative strands of DNA. Three lines above and below the grey lines represent possible reading frames with genes in blue and stop codons represented by vertical black lines. The green block upstream of htrC is its predicted sigma 70 promoter.
Key: All names read from left to right.
Blocks with thick horizontal lines: Deletion primers NohtrC, NihtrC, CihtrC and CohtrC.
Blocks with thin horizontal lines: Deletion check primers Ncheck htrC and Ccheck htrC.
Blocks with right slanting lines: Internal deletion check primers.
Blocks with left slanting lines: External sequencing primers.

Primers No-Ni htrC and Ci-Co htrC were used to amplify the region flanking *htrC*. After the gene was replaced by the *lacZ-aph* cassette the resulting mutant was verified in PCR reactions using primers N check htrC and C check htrC (figure 5.1.) with primers specific for the cassette pUCseq primer and kancass2 respectively (primer sequences are listed in Materials and Methods 2.3). The mutant gave the expected PCR product of approximately 750 bp for the 5' region and 1100 bp for the 3' region.

Figure 5.2. Schematic of PCR checks for htrC *deletion:*



During the course of the study it became necessary to verify the deletion of the mutant using different sets of primers. One set of primers was designed internal to the deleted region (figure 5.1.- right slanting lines) to give a signature PCR product of 375 bp for the intact *htrC* gene. Another set of primers flanking the *htrC* region (figure 5.1.- left slanting lines) was designed to give a PCR product of 785 bp for the intact gene and 5kb for the *htrC* gene replaced by the *lacZ aph* cassette. These PCR products are shown above in figure 5.2.

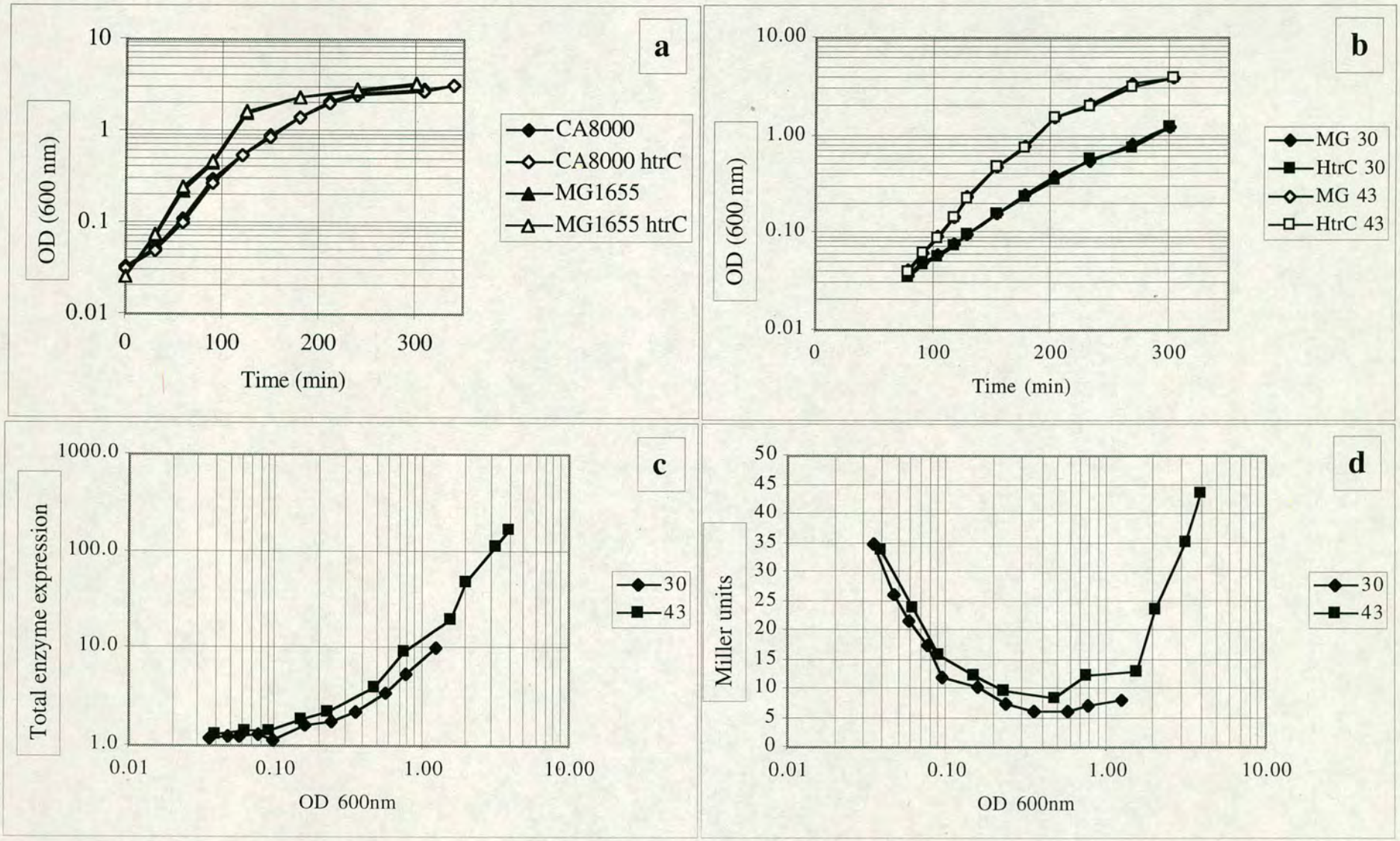
5.2.2: Phenotypic tests of the *htrC:lacZ-aph* mutant:

Growth of the *htrC* mutant was compared to that of the parent in the following sets of agar based growth conditions. Growth on LB at different temperatures (30, 37 and 45 °C), varying salt levels (800 &1200 mM), pH (5.6, 5.8, 9 and 9.2), presence of metals (zinc 2 & 3 mM, nickel 2 & 3 mM, cobalt 1 & 1.5 mM, copper 5 & 6 mM) and minimal glucose medium at different temperatures (30 and 37 °C). In all the test conditions the *htrC* mutant showed colony numbers and growth rates similar to the parent strain. Viable counts of the mutant and parent strains at 45 degrees were again very similar (9 and 14 x 10^8 ml-1 respectively).

The ability of the *htrC* mutant to grow on LB agar at 45 °C was unexpected and so the mutant was tested on LB broth at 43 °C to detect any changes in doubling times. Growth and doubling times of the *htrC* mutant and its parent MG1655 Δ*lacZ* at 43 °C were identical. To test whether the phenotype of temperature sensitivity might be a strain specific phenomenon the *htrC* deletion was transduced into strain CA8000, used by Raina et al (1990), using the *aph* resistance cassette as a selectable marker. The growth curve of the mutant CA8000 *htrC* strain was also found to be identical to its parent. Shown below in figure 5.3 (a) are the growth curves of *htrC* mutants and parent strains (MG1655 Δ*lacZ* and CA8000) in LB broth at 43 °C.

Expression of *htrC* was measured using the levels of ß galactosidase produced from the *lacZ* reporter cassette under the control of the native *htrC* promoter. Starter cultures of both MG1655 and its *htrC* mutant were grown to an optical density of 0.15 at 30 °C and were then diluted 1:4 into LB broth at either 30 or 43 °C. Growth and gene expression were monitored and the observed values are shown in figure 5.3 (b), (c) and (d).

Figure 5.3.



Key: (a) Growth curves of parental and htrC *mutant strains at 43 °C . (b) Growth curves of* MG1655 *and* htrC *mutant at 43 °C. (c) Total enzyme expression (*lacZ*) at 30 and 43 °C of* MG1655 htrC *mutant plotted against optical density. (d) Expression of* lacZ *reporter in Miller units of* MG1655 htrC mutant at 30 and 43 *°C plotted against optical density.*

Expression of *htrC* remains very low at both temperatures. There was no increase in expression of *htrC* observed upon a temperature shift from 30 to 43 °C. Gene expression in Miller units remains low in the exponential phase of growth (optical density of the culture below 1). Expression at 43 °C increases only after the culture reaches an optical density of over 1.5. This increase is not seen for the culture growing at 30 °C as a high OD was not reached during the course of the experiment.

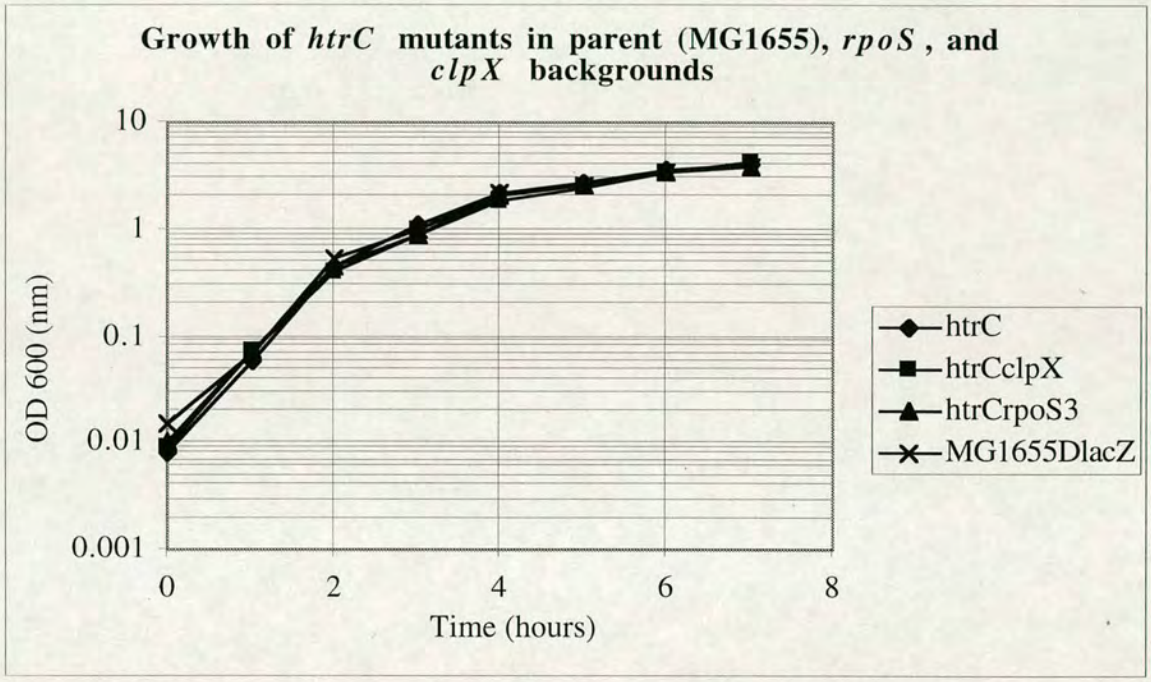
5.3. RpoS *and expression of* htrC.

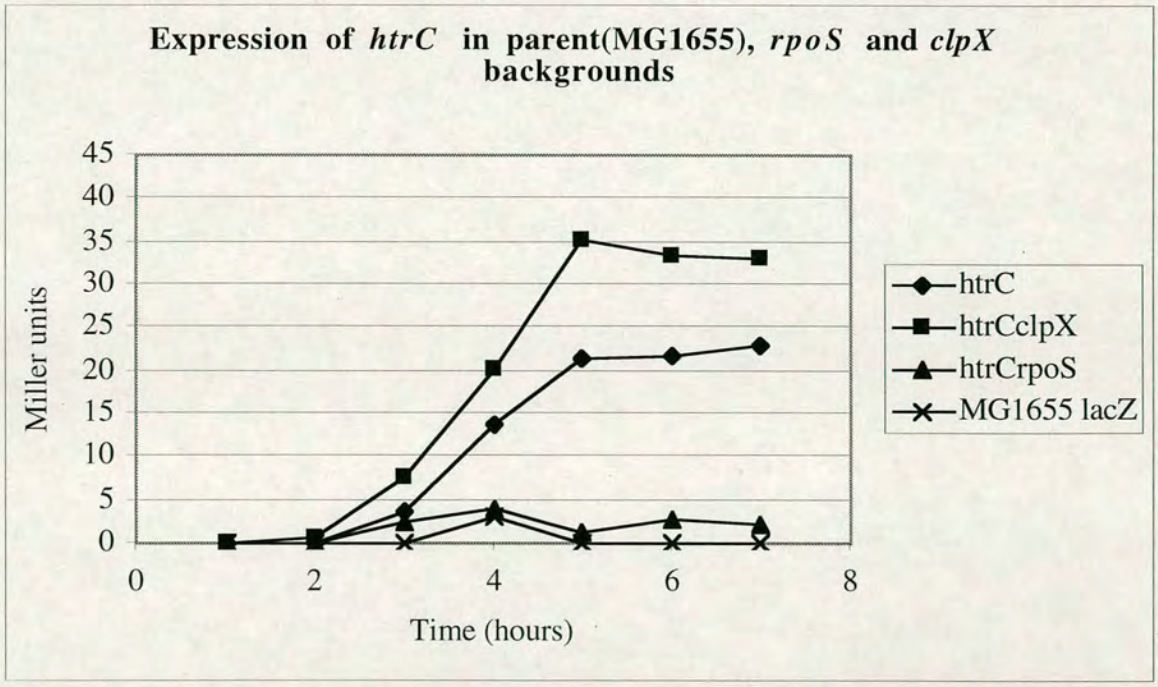
Gene expression, measured as Miller units, plotted against the optical density of the culture suggests that the expression of *htrC* increases during entry into the stationary phase of growth. However as mentioned in chapter 3 such an increase may be due to a reduction of ribosome protein synthesis in the cell during entry into stationary phase. Gene expression measured as total enzyme activity plotted against the optical density of the culture presents a different pattern of expression of *htrC*. This graph suggests that *htrC* is expressed at a lower level when the optical density of the culture is below 1. When optical density exceeds 1 *htrC* is expressed at a higher rate. Expression of *htrC* during growth in LB at 37 °C. as reported in chapter 3 (figure 3.15) is also very low (under 10 Miller units) until the culture reaches an OD of 1.5 at which time the expression increases. This increase of expression during entry into the stationary phase suggests that expression of *htrC* may be controlled by the alternative sigma factor S.

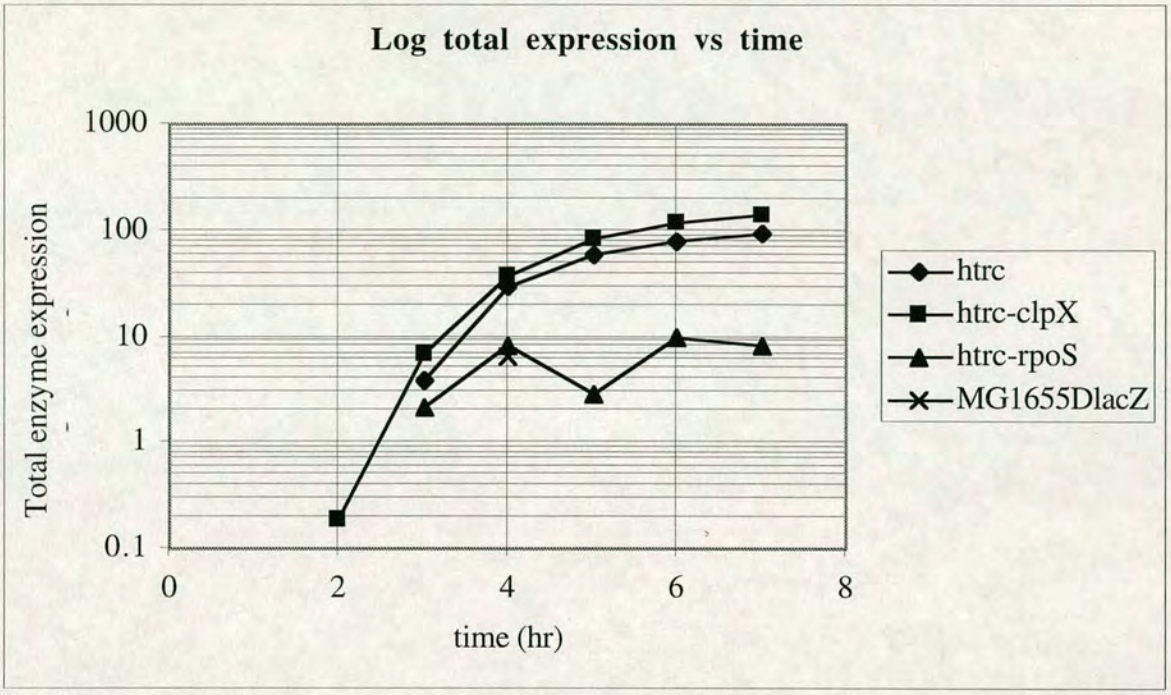
The influence of the alternative sigma factor S on the expression of *htrC* was tested by Elise Darmon (unpublished data) and is presented here with consent. Expression of *htrC* was measured in strains which were mutated for only *htrC, htrC* and *rpoS* together and *htrC* and *clpX* together. *RpoS* encodes the alternative sigma factor, sigma S, that is necessary for expression of some genes required during stationary phase (Lacour and Landini, 2004). *ClpX* encodes the protease that regulates protein levels of sigma S in the cytoplasm (Moreno et al, 2000). ß-galactosidase assays at various stages of growth show that although growth rates of all mutants are identical, expression of *htrC* in the *rpoS*

mutant is approximately 10 fold lower compared to the single *htrC* mutant. Expression of *htrC* in the *clpX* mutant is approximately 1.4 times higher which is perhaps due to the absence of the protease and its regulatory activity of sigma S. There was also no induction of expression of *htrC* observed during entry into the stationary phase of growth in the *htrC-rpoS* double mutant. Expression of *htrC* in mutants lacking *rpoS* and *clpX* is shown in figure 5.5.

Figure 5.4.







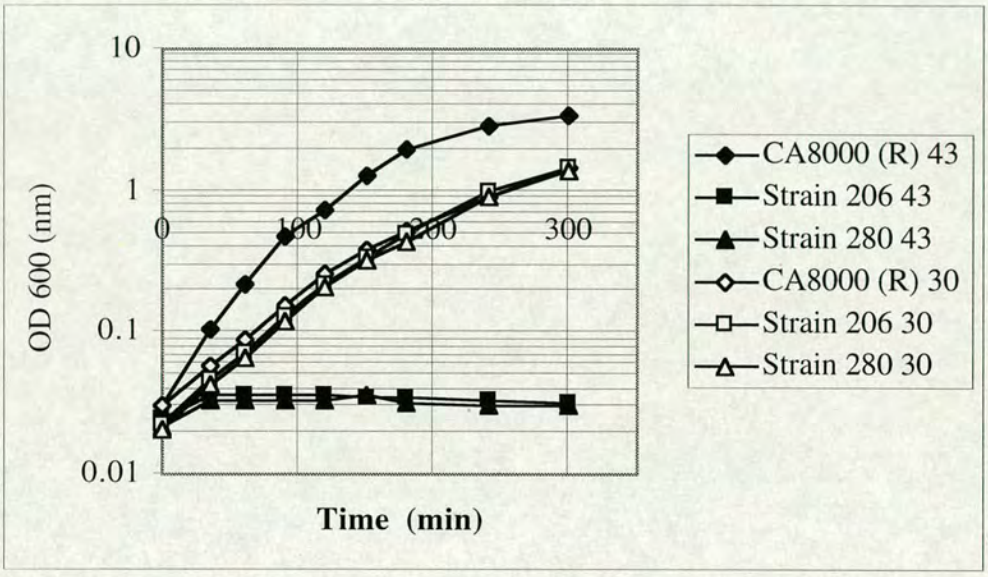
This series of experiments shows that deletion of the *htrC* did not affect the growth of the mutant at high temperatures in LB agar or broth. There was no increase in expression of *htrC* on changing its growth temperature from 30 to 43 °C. As there was no increase in expression upon temperature shift from 30 to 43 °C., it seems unlikely that the expression of *htrC* is controlled by sigma 32. Expression of *htrC* rather appears to be controlled by another alternative sigma factor (sigma S).

5.4. Temperature sensitivity of putative htrC *mutants of Dr. S. Raina.*

Since the reported temperature sensitivity of *htrC* mutants could not be reproduced in my deletion of the *htrC* gene, It was decided to study the mutants produced by the authors of the original *htrC* publication in order to determine whether strain background (i.e. additional mutations) may be responsible for the reported

phenotype. The strains I received from Dr. S. Raina were a wildtype CA8000 parent strain and two temperature sensitive, kanamycin resistant putative transposon mutants of *htrC* called 206 and 280. Compared to the CA8000 strain the mutants were indeed temperature sensitive. Shown below (figure 5.6.) is a growth curve of strains CA8000 and temperature sensitive mutants 206 and 280 carried out at 43 °C.

Figure 5.5. Growth curves of strains 206 and 280 (Dr. S. Raina) at 30 and 43 °. C.



Key: Open data points indicated growth curves at 30 °C. Closed data points indicate growth curves carried out at 43 °C. All strains were supplied by S. Raina.

This work has already shown that transducing a mutated *htrC* gene into strain CA8000 does not produce a temperature sensitive phenotype it became necessary to test whether the mutation in *htrC* or one elsewhere was causing the temperature sensitivity of mutants 206 and 280.

5.5. Transduction analysis of putative htrC mutation in strains 206, 280.

Linkage between the *htrC* gene present at 90 minutes of the *E. coli* chromosome and the temperature sensitivity of mutant strains supplied by S. Raina was tested next. This was done by testing the linkage between the tetracycline resistance on a Tn10 transposon inside *argE* (CAG12185, Nichols et al, 1998) and the reported kanamycin resistance within *htrC*. The *argE* gene is about a minute away from *htrC* on the genetic map so it would be expected that the Tn10 interrupted *argE* gene would be co-transduced with a normal copy of the *htrC* gene at a rate of 20% (making the strain kanamycin sensitive if there was indeed a kanamycin resistance gene within *htrC*). Both temperature sensitive mutants were transduced to tetracycline resistance with a lysate prepared from the *argE:Tn10* mutant. A hundred tetracycline resistant mutants were purified on LB agar tetracycline plates and were screened for kanamycin resistance. All tetracycline resistant transductants were found to be kanamycin resistant indicating that the transposon was unlikely to be within 2 minutes of *argE*.

5.6. Linkage between Tn5 and temperature sensitivity of strains 206 and 280.

To verify whether there was indeed a Tn*5* kanamycin gene giving rise to kanamycin resistance in the Ts mutants I used primers specific for the Tn*5* region flanking the kanamycin gene in a PCR reaction. This PCR gave fragments of the expected size for the kanamycin gene, which is approximately 950 base pairs for both the temperature sensitive mutants. This suggests that there is at least one Tn*5* based kanamycin resistance gene in the chromosome of the Ts mutants sent by Dr. Raina. To check if the kanamycin gene is linked to the locus causing temperature sensitivity, the Ts mutants were transduced to temperature resistance (Tr) using P1 lysates prepared from wild-type MG1655. Temperature resistant mutants at 43 °C were then checked for kanamycin resistance. All 100 Tr transductants were resistant to kanamycin demonstrating that the two phenotypes, temperature sensitivity and kanamycin resistance, are unlinked.

5.7. The htrC gene in temperature sensitive mutants 206, 280.

Since Ts and the Tn10 appeared the *htrC* gene in the temperature sensitive mutants supplied by Dr. Raina was tested for the presence of an interrupting transposon. This was done in a set of PCR reactions with primers flanking the *htrC* gene (figure 2: left slanting boxes). Chromosomal DNA of wildtype CA8000 (control), MG1655 (control), MG1655 *htrC* and Ts mutants 206 and 280 were used as templates. The CA8000 and MG1655 controls gave fragments of the expected size with the *htrC* gene intact (approximately 750 base pairs). The MG1655 *htrC* gave a PCR product of about 5 kb that includes the deleted *htrC* gene and the inserted *lacZ* and Km^R. The Ts mutants from Dr. Raina on the other hand gave PCR products of about 750 base pairs as did the wild type controls. This showed that there is indeed no *Tn*5 insertion within the *htrC* gene since the transposon itself is about 5.7 kb in length. We conclude that the Ts mutants supplied by S. Raina are not Ts because of a mutation of the *htrC* gene. The function of *htrC* remains unknown but it is clear that *htrC* plays no role in the heat shock response of *E. coli.*

5.8. Summary

Transduction experiments using a Tn*10* tet^R marker in *argE* show that the putative *htrC* mutant strains 206 and 280 supplied by Dr. Raina, although temperature sensitive, do not have temperature sensitivity linked to the *htrC* locus. Although these mutants do carry a *Tn*5 based kanamycin resistance gene, the *Tn*5 is not within the *htrC* gene itself. The temperature sensitive phenotype of the mutants sent by Dr. Raina also appears to be unlinked to the kanamycin gene on the *Tn*5 transposon. Since it is not known how many copies of *Tn*5 are on the chromosome it is impossible to say if there is a single insertion event causing the temperature sensitivity of these mutants. More importantly, with regard to *htrC*, the gene itself appears to be intact in the two temperature sensitive mutants (based on PCR analysis).

Deleting the *htrC* gene in *E. coli* strains MG1655 and CA8000 does not affect the temperature resistance of these strains. Gene expression of *htrC* measured by ß-galactosidase activity in MG1655 appears to be very low (between 30-50 Miller units). There is no upregulation of gene expression as a consequence of heat shock as would be expected from a heat shock gene. Expression of *htrC* is increased upon entry into the stationary phase of growth. Testing the expression of *htrC* in *rpoS* (encoding the stationary phase sigma factor) and *clpX* (protease that regulates activity of sigma S) mutants has shown that expression of *htrC* is stimulated by the alternative sigma factor S.

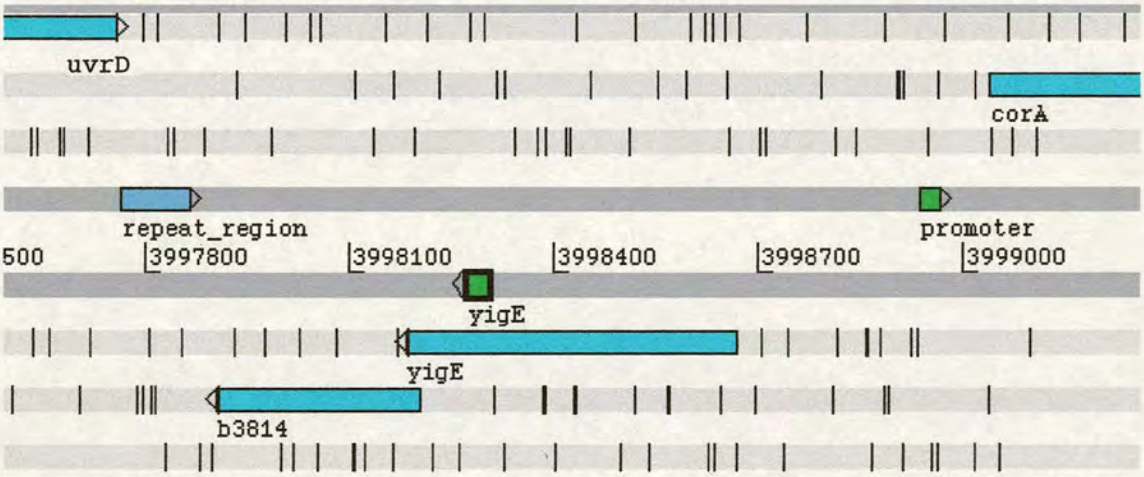
These observations lead us to conclude that *htrC* is not a heat shock gene. The deletion of *htrC* does not lead to any phenotype that we have detected. The mutant strain behaves identically to the wildtype during growth in a variety of conditions tested, showing that it plays no role in the survival of *E. coli* in these conditions. Expression of *htrC* shows no response to temperature shifts (from 30 to 43 °C). However, it is expressed (albeit at a low level) and its expression is dependent upon the stationary phase sigma factor S. Because we have not observed a heat shock related or indeed any phenotype linked to the absence of *htrC* we feel this mnemonic is both inappropriate and misleading. We therefore suggest that the gene name revert to a positional 'y' name.

Chapter 6: The *yigE* mutant and its sensitivity to Ni^{2+} and Co^{2+} ions.

6.1: Introduction.

The *yigE* mutant formed colonies at a slower rate on LB agar plates with added Ni^{2+} (3 mM) and Co^{2+} (1.5 mM). The experiments detailed in this chapter were carried out with the aim of uncovering the reason for the observed sensitivity of the *yigE* mutant to nickel and cobalt ions. The *yigE* open reading frame lies at 86.18 minute, in the *uvrD-corA* intergenic region on the *E. coli* chromosome and is predicted to encode a protein of 161 amino acids. The protein has no predicted domains and has no functions assigned to it. In *E. coli* K-12 and W3110 *yigE* has been reannotated as two separate ORFs b3814 and b3815, with b3815 being closer to *corA*. The two separate ORFs code for predicted proteins of 99 and 161 amino acids respectively. In this study b3815 (called *yigE*) was deleted but this deletion also results in a deletion of the first three codons of b3814 and its predicted sigma 70 binding site which may affect the expression and activity of the encoded protein.

Figure 6.1. The yigE *ORF and its genetic neighbours.*



Key: the figure above shows a section of the two strands (grey bars) of the K-12 MG1655 chromosome. Genes are depicted in blue and promoters in green. The three bars above and below the positive and negative strands represent the possible reading frames predicted stop codons depicted as black vertical lines.

The predicted *yigE* ORF is transcribed divergently from the gene *corA* that encodes a ubiquitous transporter of Mg^{2+}, Ni^{2+} and Co^{2+} ions. CorA is a major transporter of Mg^{2+} in *Salmonella* and other phylogenetically diverse bacterial organisms (Kehres et al, 1998). The protein can mediate influx and efflux of Mg^{2+} and this is considered to be its primary physiological function (Smith et al, 1998, Smith and Maguire, 1998). The protein can also mediate transport of Co^{2+}, Ni^{2+}, and Fe^{2+} when concentrations of these ions in the medium are cytotoxic (Hmiel et al, 1986). The position of *yigE* and phenotype of Ni^{2+} and Co^{2+} sensitivity of the *yigE* mutant suggests that *yigE* perhaps interacts with either the expression or activity of *corA*. The affects may be either at the DNA level where *yigE* might control the expression of *corA* or involve interaction at the protein level.

Interesting questions emerge about the *yigE* mutant and the gene itself. Are the intracellular levels of Ni^{2+} and Co^{2+} higher or lower in the mutant compared to the parent strain? Does *yigE* control the activity of *corA* and if it does is this at the genetic or protein level? If *yigE* is indeed a component of the *corA* import-export system and *corA* is an extremely widely distributed transporter of nickel and cobalt (Kehres et al, 1998) why then is *yigE* specific to *E. coli* and loosely conserved in *Agrobacterium* and *Rhizobium*. Experimental verification of some of these questions and the results are detailed in this chapter.

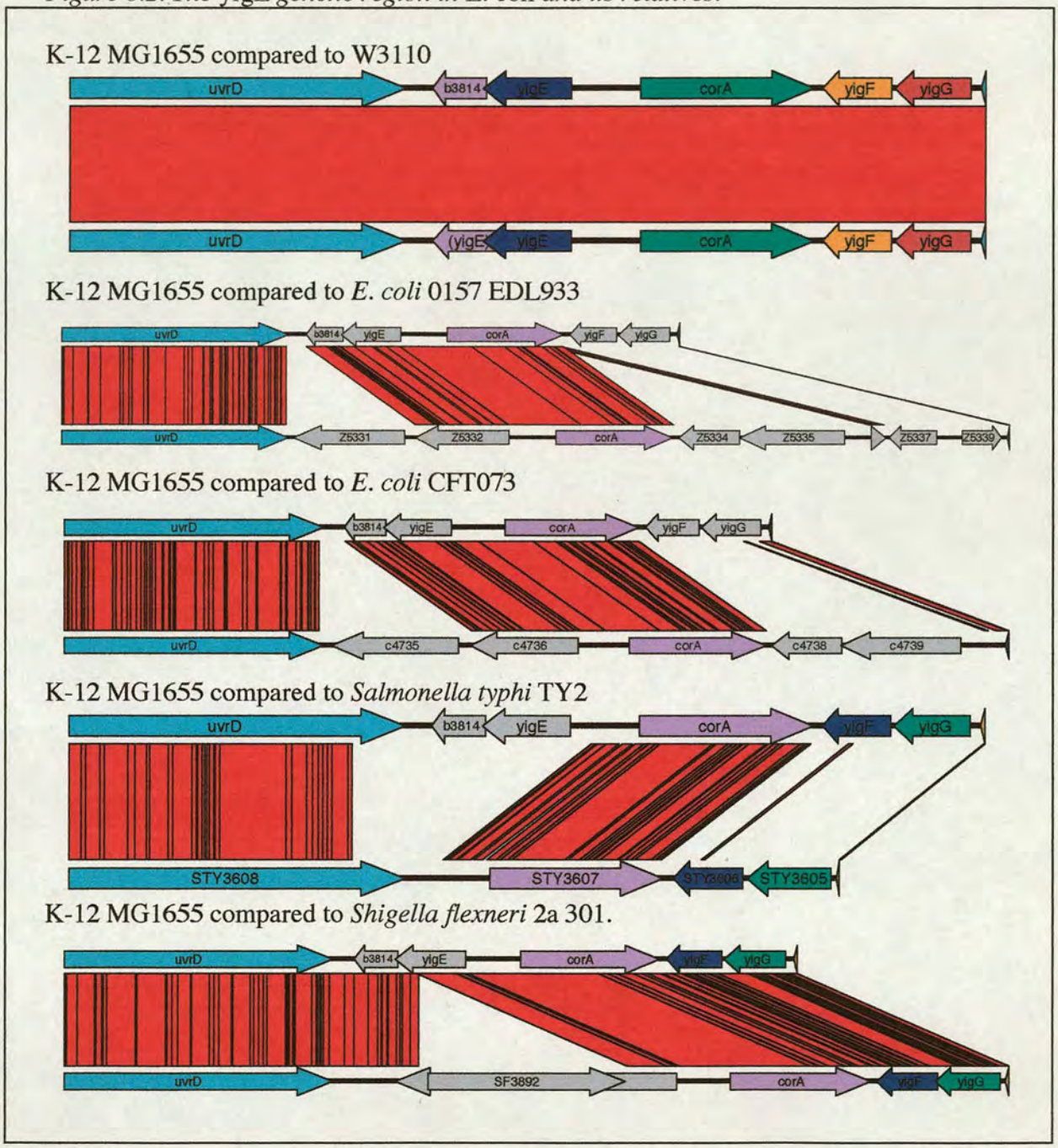
6.2: Results

6.2.1. Comparison of the *yigE* region in *E. coli* and its close relatives.

The *yigE* ORF and the genetic region flanking it is variable in strains of *E. coli* and its closest relatives *Salmonella* and *Shigella*. This is demonstrated in figure 6.2., below, which shows cartoons of the *yigE* region in MG1655 strain K-12 (upper strand)

compared to the same region in *E. coli* strains W3110, 0157, EDL933, CFT073, *Salmonella typhi* TY2 and *Shigella flexneri* 2a 301 (Lower strands).

Figure 6.2. The yigE *genetic region in* E. coli *and its relatives.*



Key: The cartoons are BLAST comparisons of the DNA regions as viewed through the online MUMMER alignment program available on Colibase

(http://colibase.bham.ac.uk). Close matches between two strands are shown in red, any base substitutions are shown as black lines and missing sequences are depicted in white.

Comparison of the *yigE* region in K-12 to that in *E. coli* 0157 EDL933 and CFT073 shows that unlike the K-12 and W3110 genomes, *yigE* shows no stop codon at position 162 and is therefore annotated as a longer ORF spanning the length of b3814 and b3815 combined. There is an additional ORF predicted between the *yigE-uvrD* intergenic region in the two pathogenic *E. coli* species that is absent in MG1655 and W3110. The b3814 and *yigE* ORFs are missing in the *Salmonella typhi* TY2 genome whereas the *Shigella flexneri* 2a 301 genome shows the full *yigE* ORF disrupted by an insertion and it is thus classified as a pseudogene. The genetic region therefore is highly variable in these closely related organisms and it is interesting to note that the *yigE* ORF is conserved in the two pathogenic *E. coli* strains but is missing or interrupted in the two closely related pathogens *S. typhi* TY2 and *S. flexneri* 2a 301.

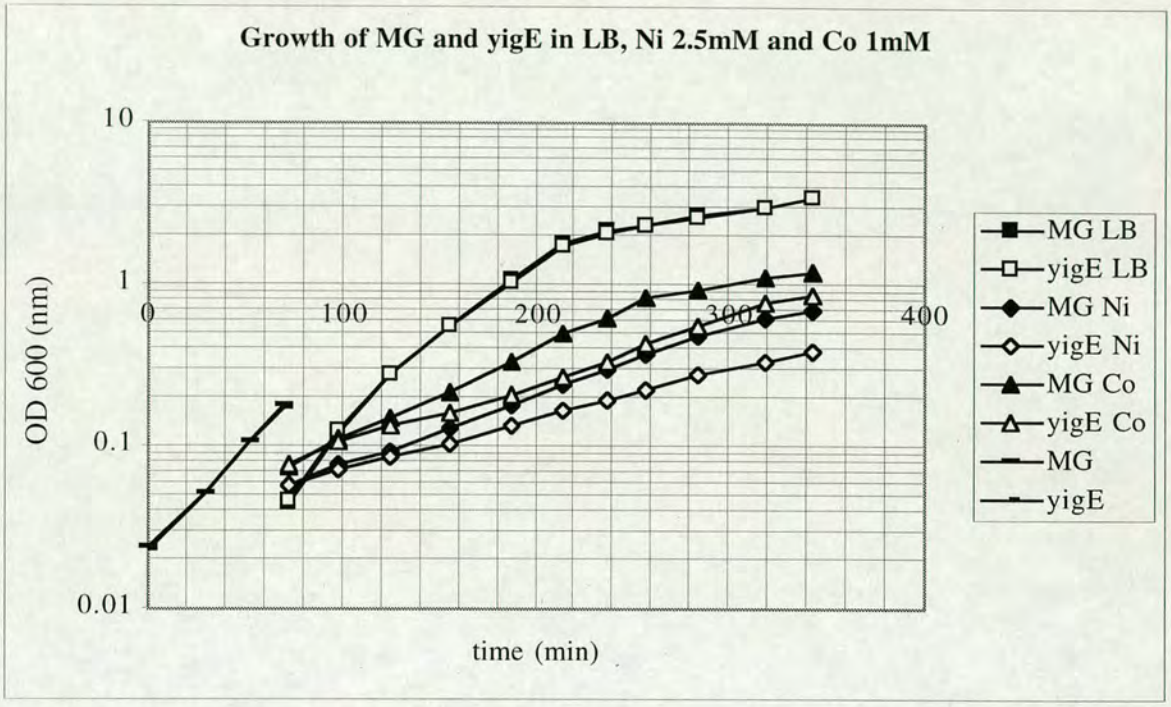
6.2.2. Growth and expression of *yigE* in the presence of Ni^{2+} and Co^{2+}.

The *yigE* mutant was tested for growth in LB broth alongside the parent MG1655 strain to quantitatively measure the slowing of growth of the *yigE* mutant in the presence of Ni^{2+} and Co^{2+}. The levels of Ni^{2+} and Co^{2+} used (2.5 mM and 1 mM) were lower than those used in plate tests (3 mM and 1.5 mM). Overnight cultures of the *yigE* mutant and MG1655 were diluted 1:100 into fresh LB medium and incubated in a shaking waterbath at 37 °C. When the two starter cultures reached on OD of 0.2 at 600nm they were diluted 1:4 into fresh LB, LB with 2.5 mM Ni^{2+} and LB with 1 mM Co^{2+} and were incubated in shaking waterbaths at 37 °C. Growth was measured at regular intervals at 600 nm.

Shown in figure 6.3. is the growth of the *yigE* mutant compared to the parent strain in the presence of Ni^{2+} and Co^{2+}. Doubling times calculated from these curves showed that the *yigE* mutant had doubling times of 96 and 119 minutes respectively in Co^{2+} and Ni^{2+}

compared to 67 and 84 minutes for the parent strain. This shows that the doubling times of the *yigE* mutant in the presence of Ni^{2+} and Co^{2+} are approximately 1.4 times longer than the parent strain.

Figure 6.3.



Key: MG1655 *and* MG1655 yigE *strains were grown to an OD of 0.2 and diluted in various test media containing either* Ni^{2+}, Co^{2+} *or plain LB. Shown above are the growth curves observed.*

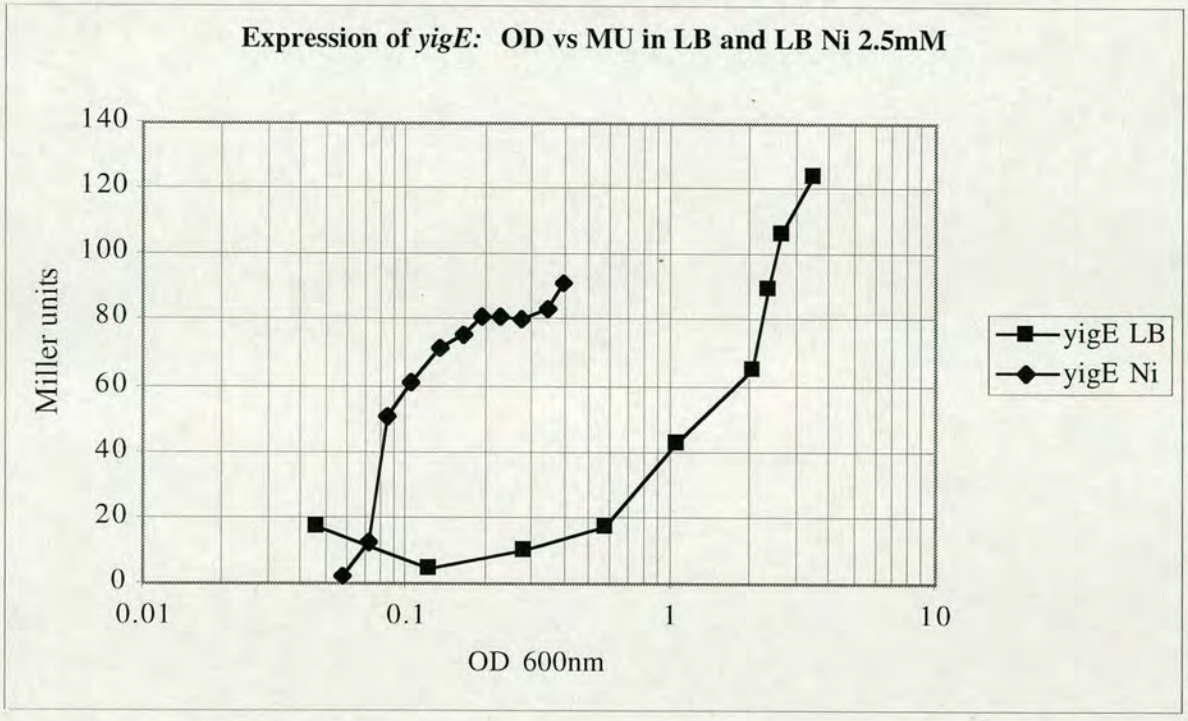
To investigate whether the expression of *yigE* is altered in the presence of nickel and cobalt ions, ß-galactosidase assays were carried out on samples collected during growth in two replicate experiments for each of these ions. The experiments were carried out as described above except that aliquots were taken for ß-galactosidase assays as well as for following growth at 600 nm. To assay expression of *yigE* in the presence of Co^{2+} it was necessary to wash cells and resuspend them in Co^{2+} free LB as the presence of Co^{2+} was

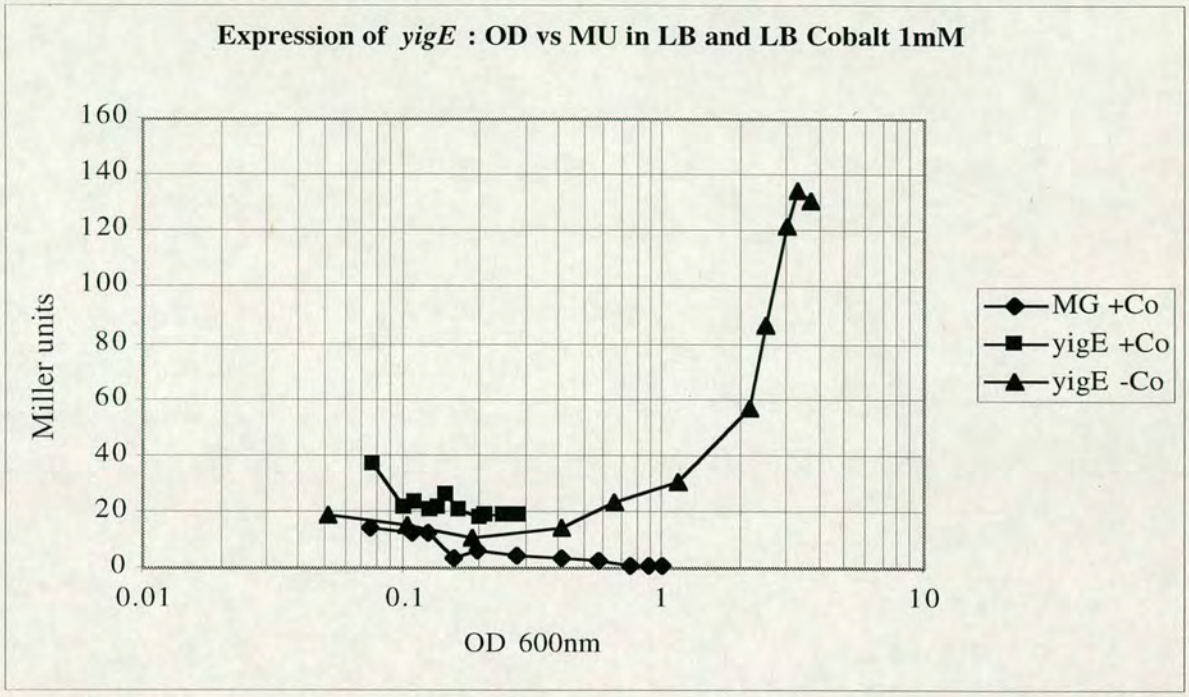
found to interfere in the detection of ONPG at 420 nm. Expression levels of *yigE* in the presence of Ni^{2+} and Co^{2+} are shown in figure 6.4. The figures show the expression in Miller units plotted against the optical density of the culture. Plotting expression against optical density of the culture was chosen over plotting expression against time since the growth rates of the two strains differ.

Expression of *yigE* in LB remained below 20 Miller units until O.D.600nm approached 1 when growth rate slowed in the approach to stationary phase. After this point the expression rate of *yigE* increased as the culture entered the stationary phase of growth (also shown in chapter 3). This suggests that expression of *yigE* may be under the control of the stationary phase specific sigma factor S, however this possibility remains untested. During exponential phase upon the addition of nickel, however, *yigE* expression showed an eight fold increase as shown in figure 6.4.

Expression of *yigE* was not similarly affected by the presence of Co^{2+}. No large increase in expression was observed when the mutant was introduced to Co^{2+} in the medium. This result suggests a difference in the expression of the *yigE* ORF in the presence of the two metal ions with nickel leading to increased expression of the ORF and cobalt not affecting the expression of *yigE*. This also eliminates the possibility that slow growth per se causes induction of *yigE* expression

Figure 6.4. Expression of yigE *in the presence of nickel and cobalt.*





Note: Expression in Miller units are plotted against the optical density of the culture as the growth rates between test conditions vary

As detailed above, a deletion of the *yigE* ORF results in a nickel and cobalt sensitivity. Presence of either ion slows growth in solid and liquid medium, however only nickel, but not cobalt, appears to induce *yigE* expression at the concentrations tested.

As noted in the Introduction, deletion of *yigE* also results in a deletion of the predicted sigma 70 site and first 9 base pairs of ORF b3814. It may be possible that the observed sensitivity to nickel and cobalt may not be solely due to a deletion of *yigE* but the effect deleting *yigE* has had on the expression of ORF b3814. The experiments listed below were carried out to test whether:

a) A plasmid borne copy of *yigE* alone could complement the deletion of the gene and restore resistance to nickel and cobalt ions.

b) 'Flipping' the FLKP2 cassette from the *yigE* deletion and leaving an in-frame deletion affects the sensitivity of the mutant to nickel and cobalt. If this restores the resistance of the mutant to nickel and cobalt then it might suggest a role played by ORF b3814.

6.2.3. Complementation of *yigE*:

To test if the *yigE* ORF when cloned in a plasmid could complement the deletion of the *yigE* ORF on the chromosome, the entire ORF was cloned onto the pBAD18-Cm vector (Guzman et al, 1995) using the *SalI-HindIII* sites within the multiple cloning site to create pBAD18-Cm-*yigE*. This places the ORF downstream of the pBAD promoter on the plasmid and thus under the control of arabinose. The recombinant plasmid and a control plasmid (pBAD18-Cm without the *yigE* ORF) were then transformed into the *yigE* mutant and the parent MG1655 strains. Serially diluted overnight cultures grown in LB with chloramphenicol were then spotted onto the surface of LB agar plates containing 0.2% arabinose with 3 mM $NiCl_2.6H_2O$ and 1.5 mM $CoCl_2$ and incubated for 48 hours. Plates were photographed and are shown in figure 6.5.

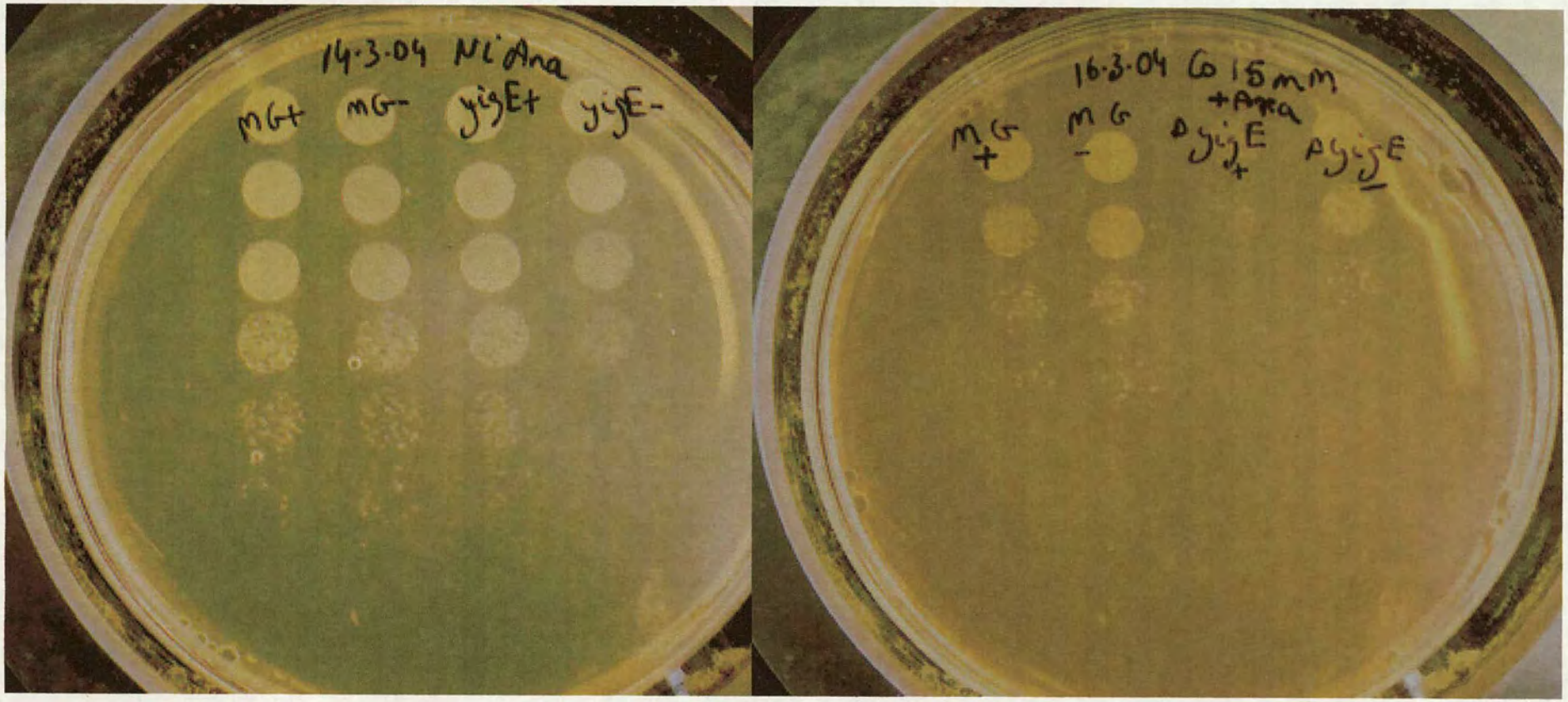
Spot and colony formation from at least two independent experiments showed that pBAD18-Cm-*yigE* restores resistance of the *yigE* mutant to 3mM $NiCl_2.6H_2O$ at a level similar to that of the parent strain. Colony and spot formation of the parent strain in the same medium appears unaffected by the plasmid either with or without the complementing *yigE* ORF.

The response in terms of growth of the *yigE* mutant with and without a complementing *yigE* ORF on LB medium with 1.5 mM $CoCl_2$ was different than to its response to nickel. The *yigE* mutant carrying pBAD18-Cm showed fewer colonies and spots that are less dense compared to the parent strain either with pBAD-18 Cm or with pBAD-18 Cm-*yigE*. However, the *yigE* mutant carrying the pBAD18-Cm-*yigE* plasmid shows

greater sensitivity to Co^{2+} than does the *yigE* mutant carrying pBAD18-Cm (Figure 6.4.). Based on these results it appears that the *yigE* ORF on plasmid pBAD18-Cm makes the *yigE* mutant even more sensitive to cobalt while restoring resistance to nickel in the same mutant

Figure 6.5: Sensitivity of MG1655 and yigE *mutant with (+) and without (-) complementing* yigE *on pBAD18-Cm* to *Ni^{2+} 3mM (left) and Co^{2+} 1.5mM (right).*

From left to right: MG+: MG1655 ΔlacZ *pBAD18-Cm-yigE, MG-: MG1655* ΔlacZ *pBAD18-Cm,* yigE+*:* ΔyigE *MG1655* ΔlacZ *pBAD18-Cm-yigE,* yigE-*:* ΔyigE *MG1655* ΔlacZ *pBAD18-Cm.*



6.2.4. In-frame deletion of *yigE* (b3815) and the role ORF b3814.

To investigate the role played by ORF b3814 in nickel and cobalt resistance of the *yigE* mutant, the FLKP2 reporter cassette was removed to leave an in-frame scar. The FLKP2 cassette was 'flipped' by introducing the yeast FLP recombinase gene on plasmid pCP20 and screening for mutants which had lost their resistance to kanamycin and chloramphenicol. The resulting mutant was tested for growth on plates of LB Ni^{2+} and Co^{2+} at concentrations of 3 and 1.5 mM respectively, with and without IPTG.

The *yigE* mutant with an intact FLKP2 cassette and the parent MG1655 Δ*lacZ* were grown alongside the 'flipped' *yigE* mutant. Overnight LB broth cultures of all three strains were serially diluted in bacterial buffer and spotted on LB agar containing Ni^{2+} or Co^{2+}. IPTG was included in a duplicate set of plates to test if expression of b3814 resulting from the *plac* promoter on the FLKP2 cassette induce enough b3814 resulting in a phenotypic effect.

Growth tests on LB agar with Ni^{2+} (3 mM) Co^{2+} (1.5 mM) with and without IPTG showed that flipping the FLKP2 cassette makes the *yigE* mutant as resistant to nickel and cobalt as the parent strain. The *yigE* mutant with FLKP2 intact remains sensitive to nickel and cobalt at the tested concentrations. There was no difference in growth observed between plates with and without IPTG suggesting that expression of b3814 from *plac* is not sufficient to restore resistance to the two metal ions tested. This experiment suggests that ORF b3814 plays a role in resistance to nickel and cobalt when the *yigE* ORF is deleted in-frame and any downstream effects of the deletion are minimised.

6.3. Discussion

Deletion-replacement of the *yigE* ORF makes the mutant sensitive to high levels of Ni^{2+} and Co^{2+} in the medium. The *yigE* mutant grows slower than the parent strain at high levels of Ni^{2+} and Co^{2+} on both solid and liquid media. Expression of *yigE* increases in the presence of Ni^{2+} but not of Co^{2+}. As the expression of *yigE* responds specifically to Ni^{2+} it may suggest that *yigE* has a role in monitoring the levels of Ni^{2+}, but not Co^{2+}, in the growth medium. Overexpressing the *yigE* ORF on a plasmid complements the Ni^{2+} sensitivity phenotype, however, overexpression of *yigE* in the presence of Co^{2+} makes the mutant hypersensitive to Co^{2+} ions. This may perhaps explain the difference in expression patterns of *yigE* in the presence of Ni^{2+} (high) and Co^{2+} (low).

The gene downstream of *yigE*, b3814, has also been observed to play a role in Ni^{2+} and Co^{2+} resistance as creating an in-frame deletion of *yigE* results in parental levels of Ni^{2+} and Co^{2+} resistance in the *yigE* mutant. This may mean that the *yigE* phenotype is perhaps due to a frameshift caused by the insertion of the large FLKP2 cassette. But having shown that complementing *yigE* on a plasmid restores Ni^{2+} resistance in the mutant, it is not possible to discount the role played by *yigE*. Perhaps the two ORFs form an operon whose expression is initiated upstream of *yigE* and both play a role in resistance to nickel and cobalt ions.

There are numerous experiments that could not be carried out here due to time constraints. One of the most essential experiments that needs to be performed is to delete the two ORFs – b3814 and *yigE* singly and together and test the resulting mutants for growth on Ni^{2+} and Co^{2+} containing media. This would help elucidate the role played by the two ORFs together or individually in response to growth in the presence of the two metal ions. It would also be interesting to test expression patterns of b3814 in response to Ni^{2+} and Co^{2+} and compare them to those reported for *yigE* here. Primer extension experiments also need to be performed to determine the transcriptional origins of the two genes and determine if the two genes form an operon.

The close proximity of *yigE* to the *corA* Mg^{2+}, Ni^{2+}, Co^{2+} transporter suggests that *yigE* may interact with *corA* at the genetic or protein level and perhaps influences the transport of metal ions across the membrane. CorA can transport Ni^{2+} and Co^{2+} across the cell membrane and inactivation of *corA* confers resistance to metal mediated toxicity in *E. coli* and *Salmonella* (Park et al, 1976; Hmiel et al, 1989). Measuring intracellular levels of Mg^{2+}, Ni^{2+} and Co^{2+} during growth of the combined mutant and its parent in high and low concentrations of these metals would help understand the biological significance of the observed sensitivity. To understand any genetic interactions between *yigE*/b3814 and *corA,* it would be interesting to compare the expression patterns of *corA* in the presence of heavy metal ions in a wildtype strain and combined *yigE*/b3814 mutant. The two genes may function as negative regulators of *corA* expression during growth in the presence of Ni^{2+} and Co^{2+} and their deletion may result in increased import or decreased export of these toxic metal ions.

If a deletion of the two ORFs does not produce an observable change in *corA* expression it would suggest that they may interact with CorA at the protein level. Disrupting the transport function of the CorA protein by including Mg^{2+}, cation hexamides or sodium azide in the growth medium (Chamnongpol and Groisman, 2002) in the presence of Ni^{2+} and Co^{2+} and studying the growth of the *yigE*/b3814 mutant alongside its parent strain would be one way to test any interactions between the products of the two ORFs and CorA.

Understanding the function of this operon is important since the genes, although specific to *E. coli,* may function as novel regulators of the ubiquitous heavy metal transport protein *corA*. The *corA* transporter is ubiquitously distributed in nature (Kehres et al, 1998) and has been shown to be involved in maintaining Mg^{2+} homeostasis in *Salmonella* (Chamnongpol and Groisman, 2002). Maintaining Mg^{2+} homeostasis is important since the metal ion is essential for a wide variety of physiological functions. Studies in *E. coli* and *Saccharomyces cerevisiae* have shown that cytoplasmic

concentrations of Mg^{2+} change less than two and four fold respectively when environmental levels change a 1000 fold (Silver and Clark, 1971; Graschopf et al, 2001). Uptake of Mg^{2+} mediated by CorA has been demonstrated to be essential for viability of the gastric pathogen *Helicobacter pylori* and is implicated in adaptation to low- Mg^{2+} conditions predominant in the gastric environment (Pfeiffer et al, 2002).

There is however little known about the regulation of *corA*. In the yeast *S. cerevisiae,* high Mg^{2+} concentrations leads to degradation of the *corA* mRNA and protein (Graschopf et al, 2001). In contrast, *Salmonella* shows no alterations in transcription of *corA* in response to changes in Mg^{2+} concentrations (Snavely et al, 1991, Chamnongpol and Groisman, 2002). *CorA* therefore has evolved different mechanisms or patterns of regulation in different organisms. As noted above, the region flanking *yigE* is variable in *E. coli* and its closely related genomes. In the three *E. coli* genomes *yigE* is annotated as either two distinct ORFs b3814/*yigE* (K-12 genome) or a single ORF spanning the length of two (0157 and CFT073). In the *Salmonella* and *Shigella* genomes *yigE* is either absent or interrupted, respectively. This suggests that if *yigE* does indeed modulate the expression or activity of *corA* it is perhaps a mechanism that is specific to the *E. coli* genomes.

Understanding the cause of the observed sensitivity of the *yigE* mutants to nickel and cobalt would be interesting even if it is later discovered that *yigE* and its gene products do not interact with *corA*. The suggested experiments would help uncover any relationship that may exist between b3814/*yigE* and *corA* and their products or could go to show that b3814/*yigE* products can mediate transport of the nickel and cobalt independent of CorA.

References

Abergel, C., Monchois, V., Chenivesse, S., Jeudy, S., and Claverie, J.M., Crystallization and preliminary crystallographic study of b0220, an `ORFan' protein of unknown function from *Escherichia coli*. Acta Cryst. 2000. D56, 1694-1695.

Abou-Jaoude A., Pascal M.C., Casse F., Chippaux M., Isolation and phenotypes of mutants from *Escherichia coli* K-12 defective in nitrite-reductase activity. FEMS Microbiol. Let. 1978 (3), 235-239.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.
Basic local alignment search tool.
J Mol Biol. 1990 Oct 5;215(3):403-10.

Alimi JP, Poirot O, Lopez F, Claverie JM. Reverse transcriptase-polymerase chain reaction validation of 25 "orphan" genes from *Escherichia coli* K-12 MG1655.
Genome Res. 2000 Jul;10(7):959-66.

Arfin SM, Long AD, Ito ET, Tolleri L, Riehle MM, Paegle ES, Hatfield GW.
Global gene expression profiling in *Escherichia coli* K12. The effects of integration host factor.
J Biol Chem. 2000 Sep 22;275(38):29672-84.

Arigoni F, Talabot F, Peitsch M, Edgerton MD, Meldrum E, Allet E, Fish R, Jamotte T, Curchod ML, Loferer H. A genome-based approach for the identification of essential bacterial genes. Nat Biotechnol. 1998 Sep;16(9):851-6.

Badarinarayana V, Estėp PW 3rd, Shendure J, Edwards J, Tavazoie S, Lam F, Church GM. Selection analyses of insertional mutants using subgenic-resolution arrays.
Nat Biotechnol. 2001 Nov;19(11):1060-5.

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP, Jovanovich SB, Gates CM, Feldman RA, Spudich JL, Spudich EN, DeLong EF.
Bacterial rhodopsin: evidence for a new type of phototrophy in the sea.
Science. 2000 Sep 15;289(5486):1902-6.

Bergthorsson U, Ochman H.
Distribution of chromosome length variation in natural isolates of *Escherichia coli*.
Mol Biol Evol. 1998 Jan;15(1):6-16.

Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN.
Global analysis of mRNA decay and abundance in *Escherichia coli* at single-gene resolution using two-color fluorescent DNA microarrays.
Proc Natl Acad Sci U S A. 2002 Jul 23;99(15):9697-702.

Bhagwat AA, Bhagwat M.
Comparative analysis of transcriptional regulatory elements of glutamate-dependent acid-resistance systems of Shigella flexneri and *Escherichia coli* O157:H7.
FEMS Microbiol Lett. 2004 May 1;234(1):139-47.

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y.
The complete genome sequence of *Escherichia coli* K-12.
Science. 1997 Sep 5;277(5331):1453-74.

Bochner BR. New technologies to assess genotype-phenotype relationships.
Nat Rev Genet. 2003 Apr;4(4):309-14.

Boucher Y, Nesbo CL, Doolittle WF.

Microbial genomes: dealing with diversity.
Curr Opin Microbiol. 2001 Jun;4(3):285-9.

Brokx SJ, Ellison M, Locke T, Bottorff D, Frost L, Weiner JH.
Genome-wide analysis of lipoprotein expression in *Escherichia coli* MG1655.
J Bacteriol. 2004 May;186(10):3254-8.

Brown TD, Jones-Mortimer MC, Kornberg HL.
The enzymic interconversion of acetate and acetyl-coenzyme A in *Escherichia coli*.
J Gen Microbiol. 1977 Oct;102(2):327-36.

Chain PS, Carniel E, Larimer FW, Lamerdin J, Stoutland PO, Regala WM, Georgescu AM, Vergez LM, Land ML, Motin VL, Brubaker RR, Fowler J, Hinnebusch J, Marceau M, Medigue C, Simonet M, Chenal-Francisque V, Souza B, Dacheux D, Elliott JM, Derbise A, Hauser LJ, Garcia E.
Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis.
Proc Natl Acad Sci U S A. 2004 Sep 21;101(38):13826-31. Epub 2004 Sep 09.

Chamnongpol S, Groisman EA.
Mg2+ homeostasis and avoidance of metal toxicity.
Mol Microbiol. 2002 Apr;44(2):561-71.

Charlebois, R.L. Clarke, P.G.D., Beiko, R.G., St. Jean, A..
Characterization of species-specific genes using a flexible, web-based querying system.
FEMS Microbiology Letters 225 (2003) 213-220.

Cheung KJ, Badarinarayana V, Selinger DW, Janse D, Church GM.
A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*.

Genome Res. 2003 Feb;13(2):206-15.

Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG.
Massive gene decay in the leprosy bacillus.
Nature. 2001 Feb 22;409(6823):1007-11.

Dahan S, Knutton S, Shaw RK, Crepin VF, Dougan G, Frankel G.
Transcriptome of enterohemorrhagic *Escherichia coli* O157 adhering to eukaryotic plasma membranes.
Infect Immun. 2004 Sep;72(9):5452-9.

Datsenko KA, Wanner BL.
One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products.
Proc Natl Acad Sci U S A. 2000 Jun 6;97(12):6640-5.

Day WA Jr, Fernandez RE, Maurelli AT.
Pathoadaptive mutations that enhance virulence: genetic organization of the cadA regions of Shigella spp.
Infect Immun. 2001 Dec;69(12):7471-80.

DeLisa MP, Wu CF, Wang L, Valdes JJ, Bentley WE.
DNA microarray-based identification of genes controlled by autoinducer 2-stimulated quorum sensing in *Escherichia coli*.
J Bacteriol. 2001 Sep;183(18):5239-47.

Dharmadi Y, Gonzalez R.
DNA microarrays: experimental issues, data analysis, and application to bacterial systems.
Biotechnol Prog. 2004 Sep-Oct;20(5):1309-24.

Dobrindt U, Agerer F, Michaelis K, Janka A, Buchrieser C, Samuelson M, Svanborg C, Gottschalk G, Karch H, Hacker J.
Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays.
J Bacteriol. 2003 Mar;185(6):1831-40.

dos Reis M, Wernisch L, Savva R.
Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome.
Nucleic Acids Res. 2003 Dec 1;31(23):6976-85.

Edwards JS, Palsson BO.
The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities.
Proc Natl Acad Sci U S A. 2000 May 9;97(10):5528-33.

Fernandez De Henestrosa AR, Ogi T, Aoyagi S, Chafin D, Hayes JJ, Ohmori H, Woodgate R. Identification of additional genes belonging to the LexA regulon in *Escherichia coli.*
Mol Microbiol. 2000 Mar;35(6):1560-72.

Eguchi Y, Oshima T, Mori H, Aono R, Yamamoto K, Ishihama A, Utsumi R.
Transcriptional regulation of drug efflux genes by EvgAS, a two-component system in *Escherichia coli.*

Microbiology. 2003 Oct;149(Pt 10):2819-28.

Fountoulakis M, Takacs MF, Berndt P, Langen H, Takacs B.
Enrichment of low abundance proteins of *Escherichia coli* by hydroxyapatite chromatography.
Electrophoresis. 1999 Aug;20(11):2181-95.

Fountoulakis M, Gasser R.
Proteomic analysis of the cell envelope fraction of *Escherichia coli*.
Amino Acids. 2003;24(1-2):19-41.

Freiberg C, Wieland B, Spaltmann F, Ehlert K, Brotz H, Labischinski H.
Identification of novel essential *Escherichia coli* genes conserved among pathogenic bacteria.
J Mol Microbiol Biotechnol. 2001 Jul;3(3):483-9.

Fukiya S, Mizoguchi H, Tobe T, Mori H.
Extensive genomic diversity in pathogenic *Escherichia coli* and Shigella Strains revealed by comparative genomic hybridization microarray.
J Bacteriol. 2004 Jun;186(12):3911-21.

Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, Kapatral V, D'Souza M, Baev MV, Grechkin Y, Mseeh F, Fonstein MY, Overbeek R, Barabasi AL, Oltvai ZN, Osterman AL.
Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655.
J Bacteriol. 2003 Oct;185(19):5673-84.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M.

Functional profiling of the Saccharomyces cerevisiae genome.

Nature. 2002 Jul 25;418(6896):387-91.

Gong S, Richard H, Foster JW.

YjdE (AdiC) is the arginine:agmatine antiporter essential for arginine-dependent acid resistance in *Escherichia coli*.

J Bacteriol. 2003 Aug;185(15):4402-9.

Graschopf A, Stadler JA, Hoellerer MK, Eder S, Sieghardt M, Kohlwein SD, Schweyen RJ.

The yeast plasma membrane protein Alr1 controls Mg2+ homeostasis and is subject to Mg2+-dependent control of its synthesis and degradation.

J Biol Chem. 2001 May 11;276(19):16216-22. Epub 2001 Feb 20.

Grimm V, Ezaki S, Susa M, Knabbe C, Schmid RD, Bachmann TT.

Use of DNA microarrays for rapid genotyping of TEM beta-lactamases that confer resistance.

J Clin Microbiol. 2004 Aug;42(8):3766-74.

Gustafsson P, Nordstrom K, Normark S.
Outer penetration barrier of *Escherichia coli* K-12: kinetics of the uptake of gentian violet by wild type and envelope mutants.
J Bacteriol. 1973 Nov;116(2):893-900.

Hagiwara D, Yamashino T, Mizuno T.
A Genome-Wide View of the *Escherichia coli* BasS-BasR Two-component System Implicated in Iron-responses.
Biosci Biotechnol Biochem. 2004 Aug;68(8):1758-67.

Hamilton CM, Aldea M, Washburn BK, Babitzke P, Kushner SR.
New method for generating deletions and gene replacements in *Escherichia coli.*
J Bacteriol. 1989 Sep;171(9):4617-22.

Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H.
Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12.
DNA Res. 2001 Feb 28;8(1):11-22. Erratum in: DNA Res 2001 Apr 27;8(2):96.

Herring CD, Blattner FR.
Conditional lethal amber mutations in essential *Escherichia coli* genes.
J Bacteriol. 2004 May;186(9):2673-81.

Herring CD, Glasner JD, Blattner FR.
Gene replacement without selection: regulated suppression of amber mutations in *Escherichia coli.*
Gene. 2003 Jun 5;311:153-63.

Hmiel SP, Snavely MD, Miller CG, Maguire ME. Magnesium transport in Salmonella typhimurium: characterization of magnesium influx and cloning of a transport gene. J Bacteriol. 1986 Dec;168(3):1444-50.

Hommais F, Krin E, Laurent-Winter C, Soutourina O, Malpertuy A, Le Caer JP, Danchin A, Bertin P.
Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS.
Mol Microbiol. 2001 Apr;40(1):20-36.

Hua Q, Yang C, Oshima T, Mori H, Shimizu K.
Analysis of gene expression in *Escherichia coli* in response to changes of growth-limiting nutrient in chemostat cultures.
Appl Environ Microbiol. 2004 Apr;70(4):2354-66.

Hung SP, Baldi P, Hatfield GW.
Global gene expression profiling in *Escherichia coli* K12. The effects of leucine-responsive regulatory protein.
J Biol Chem. 2002 Oct 25;277(43):40309-23. Epub 2002 Jul 18.

Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC.
Global transposon mutagenesis and a minimal Mycoplasma genome.
Science. 1999 Dec 10;286(5447):2165-9.

Ibanez-Ruiz M, Robbe-Saule V, Hermant D, Labrude S, Norel F. Identification of RpoS (sigma(S))-regulated genes in Salmonella enterica serovar typhimurium.
J Bacteriol. 2000 Oct;182(20):5749-56.

Inoue K, Chen J, Kato I, Inouye M.

Specific growth inhibition by acetate of an *Escherichia coli* strain expressing Era-dE, a dominant negative Era mutant.
J Mol Microbiol Biotechnol. 2002 Jul;4(4):379-88.

Ishii A, Oshima T, Sato T, Nakasone K, Mori H, Kato C.
Analysis of hydrostatic pressure effects on transcription in *Escherichia coli* by DNA microarray procedure.
Extremophiles. 2004 Aug 31 [Epub ahead of print]

Itoh T, Okayama T, Hashimoto H, Takeda J, Davis RW, Mori H, Gojobori T.
A low rate of nucleotide changes in *Escherichia coli* K-12 estimated from a comparison of the genome sequences between two different substrains. FEBS Lett. 1999, 450(1-2), 72-76.

Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, Zhang X, Zhang J, Yang G, Wu H, Qu D, Dong J, Sun L, Xue Y, Zhao A, Gao Y, Zhu J, Kan B, Ding K, Chen S, Cheng H, Yao Z, He B, Chen R, Ma D, Qiang B, Wen Y, Hou Y, Yu J.
Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157.
Nucleic Acids Res. 2002 Oct 15;30(20):4432-41.

Jasin M, Schimmel P.
Deletion of an essential gene in *Escherichia coli* by site-specific recombination with linear DNA fragments.
J Bacteriol. 1984 Aug;159(2):783-6.

Kabir MS, Sagara T, Oshima T, Kawagoe Y, Mori H, Tsunedomi R, Yamada M.
Effects of mutations in the rpoS gene on cell viability and global gene expression under nitrogen starvation in *Escherichia coli*.
Microbiology. 2004 Aug;150(Pt 8):2543-53.

Kakuda H, Hosono K, Shiroishi K, Ichihara S.
Identification and characterization of the ackA (acetate kinase A)-pta (phosphotransacetylase) operon and complementation analysis of acetate utilization by an ackA-pta deletion mutant of *Escherichia coli.*
J Biochem (Tokyo). 1994 Oct;116(4):916-22.

Kauffman KJ, Prakash P, Edwards JS.
Advances in flux balance analysis.
Curr Opin Biotechnol. 2003 Oct;14(5):491-6.

Kehres DG, Lawyer CH, Maguire ME. The CorA magnesium transporter gene family.
Microb Comp Genomics. 1998;3(3):151-69.

Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, Yanofsky C.
DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli.*
Proc Natl Acad Sci U S A. 2000 Oct 24;97(22):12170-5.

Kim SK.
Http://C. elegans: mining the functional genomic landscape.
Nat Rev Genet. 2001 Sep;2(9):681-9. Review.

Kleckner N, Bender J, Gottesman S. Uses of transposons with emphasis on Tn10.
Methods Enzymol. 1991;204:139-80. Review.

Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, Boland F, Brignell SC, Bron S, Bunai K, Chapuis J, Christiansen LC, Danchin A, Debarbouille M, Dervyn E, Deuerling E, Devine K, Devine SK, Dreesen O, Errington J, Fillinger S, Foster SJ, Fujita Y, Galizzi A, Gardan

R, Eschevins C, Fukushima T, Haga K, Harwood CR, Hecker M, Hosoya D, Hullo MF, Kakeshita H, Karamata D, Kasahara Y, Kawamura F, Koga K, Koski P, Kuwana R, Imamura D, Ishimaru M, Ishikawa S, Ishio I, Le Coq D, Masson A, Mauel C, Meima R, Mellado RP, Moir A, Moriya S, Nagakawa E, Nanamiya H, Nakai S, Nygaard P, Ogura M, Ohanan T, O'Reilly M, O'Rourke M, Pragai Z, Pooley HM, Rapoport G, Rawlins JP, Rivas LA, Rivolta C, Sadaie A, Sadaie Y, Sarvas M, Sato T, Saxild HH, Scanlan E, Schumann W, Seegers JF, Sekiguchi J, Sekowska A, Seror SJ, Simon M, Stragier P, Studer R, Takamatsu H, Tanaka T, Takeuchi M, Thomaides HB, Vagner V, van Dijl JM, Watabe K, Wipat A, Yamamoto H, Yamamoto M, Yamamoto Y, Yamane K, Yata K, Yoshida K, Yoshikawa H, Zuber U, Ogasawara N.
Essential Bacillus subtilis genes.
Proc Natl Acad Sci U S A. 2003 Apr 15;100(8):4678-83. Epub 2003 Apr 07.

Kumari S, Tishel R, Eisenbach M, Wolfe AJ.
Cloning, characterization, and functional expression of acs, the gene which encodes acetyl coenzyme A synthetase in *Escherichia coli*.
J Bacteriol. 1995 May;177(10):2878-86.

Lacour S, Landini P.
SigmaS-dependent gene expression at the onset of stationary phase in *Escherichia coli*: function of sigmaS-dependent genes and identification of their promoter sequences.
J Bacteriol. 2004 Nov;186(21):7186-95.

Lai EM, Nair U, Phadke ND, Maddock JR. Proteomic screening and identification of differentially distributed membrane proteins in *Escherichia coli*.
Mol Microbiol. 2004 May;52(4):1029-44.

Lan R, Reeves PR.
Intraspecies variation in bacterial genomes: the need for a species genome concept.
Trends Microbiol. 2000 Sep;8(9):396-401. Review.

Lehnen D, Blumer C, Polen T, Wackwitz B, Wendisch VF, Unden G.
LrhA as a new transcriptional key regulator of flagella, motility and chemotaxis genes in *Escherichia coli*.
Mol Microbiol. 2002 Jul;45(2):521-32.

LeVine SM, Ardeshir F, Ames GF. Isolation and Characterization of acetate kinase and phosphotransacetylase mutants of *Escherichia coli* and Salmonella typhimurium.
J Bacteriol. 1980 Aug;143(2):1081-5.

Link AJ, Phillips D, Church GM.
Methods for generating precise deletions and insertions in the genome of wild-type *Escherichia coli*: application to open reading frame characterization.
J Bacteriol. 1997 Oct;179(20):6228-37.

Liu X, De Wulf P.
Probing the ArcA-P modulon of *Escherichia coli* by whole genome transcriptional analysis and sequence recognition profiling.
J Biol Chem. 2004 Mar 26;279(13):12588-97. Epub 2004 Jan 07.

Lobner-Olesen A, Marinus MG, Hansen FG.
Role of SeqA and Dam in *Escherichia coli* gene expression: a global/microarray analysis.
Proc Natl Acad Sci U S A. 2003 Apr 15;100(8):4672-7. Epub 2003 Apr 07.

MacBeath G, Schreiber SL.
Printing proteins as microarrays for high-throughput function determination.
Science. 2000 Sep 8;289(5485):1760-3.

Manna D, Breier AM, Higgins NP.

Microarray analysis of transposition targets in *Escherichia coli*: the impact of transcription.
Proc Natl Acad Sci U S A. 2004 Jun 29;101(26):9780-5. Epub 2004 Jun 21.

Martin RG, Rosner JL.
Genomics of the marA/soxS/rob regulon of *Escherichia coli*: identification of directly activated promoters by application of molecular genetics and informatics to microarray data.
Mol Microbiol. 2002 Jun;44(6):1611-24.

Masters M. The frequency of P1 transduction of the genes of *Escherichia coli* as a function of chromosomal position: preferential transduction of the origin of replication.
Mol Gen Genet. 1977 Oct 20;155(2):197-202.

Masuda N, Church GM. *Escherichia coli* gene expression responsive to levels of the response regulator EvgA.
J Bacteriol. 2002 Nov;184(22):6225-34.

Masuda N, Church GM.
Regulatory network of acid resistance genes in *Escherichia coli*.
Mol Microbiol. 2003 May;48(3):699-712.

Merlin C, McAteer S, Masters M.
Tools for characterization of *Escherichia coli* genes of unknown function.
J Bacteriol. 2002 Aug;184(16):4573-81.

Minagawa S, Ogasawara H, Kato A, Yamamoto K, Eguchi Y, Oshima T, Mori H, Ishihama A, Utsumi R.
Identification and molecular characterization of the Mg2+ stimulon of *Escherichia coli*.
J Bacteriol. 2003 Jul;185(13):3696-702.

Monchois, V., Aberge, C., Sturgis, J., Jeudy, S. and Claverie, J.M.,
Escherichia coli ykfE ORFan Gene Encodes a Potent Inhibitor of C-type Lysozyme
J. Biol. Chem. (2001) Vol. 276, No. 21, May 25, pp. 18437–18441.

Moreno M, Audia JP, Bearson SM, Webb C, Foster JW.
Regulation of sigma S degradation in Salmonella enterica var typhimurium: in vivo interactions between sigma S, the response regulator MviA(RssB) and ClpX.
J Mol Microbiol Biotechnol. 2000 Apr;2(2):245-54.

Mori H, Isono K, Horiuchi T, Miki T.
Functional genomics of *Escherichia coli* in Japan.
Res Microbiol. 2000 Mar;151(2):121-8. Review.

Myung K, Kolodner RD. Induction of genome instability by DNA damage in Saccharomyces cerevisiae.
DNA Repair (Amst). 2003 Mar 1;2(3):243-58.

Naylor SW, Low JC, Besser TE, Mahajan A, Gunn GJ, Pearce MC, McKendrick IJ, Smith DG, Gally DL.
Lymphoid follicle-dense mucosa at the terminal rectum is the principal site of colonization of enterohemorrhagic *Escherichia coli* O157:H7 in the bovine host.
Infect Immun. 2003 Mar;71(3):1505-12.

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, Fraser CM, et al.
Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima.

Nature. 1999 May 27;399(6734):323-9.

Newman BJ, Masters M. The variation in frequency with which markers are transduced by phage P1 is primarily a result of discrimination during recombination.
Mol Gen Genet. 1980;180(3):585-9.

Nichols BP, Shafiq O, Meiners V. Sequence analysis of Tn10 insertion sites in a collection of *Escherichia coli* strains used for genetic mapping and strain construction.
J Bacteriol. 1998 Dec;180(23):6408-11.

Ochman H, Lawrence JG, Groisman EA.
Lateral gene transfer and the nature of bacterial innovation.
Nature. 2000 May 18;405(6784):299-304. Review.

Oh MK, Liao JC.
DNA microarray detection of metabolic responses to protein overproduction in *Escherichia coli.*
Metab Eng. 2000 Jul;2(3):201-9.

Oh MK, Rohlin L, Kao KC, Liao JC.
Global expression profiling of acetate-grown *Escherichia coli.*
J Biol Chem. 2002 Apr 12;277(15):13175-83. Epub 2002 Jan 28.

Oshima T, Wada C, Kawagoe Y, Ara T, Maeda M, Masuda Y, Hiraga S, Mori H.
Genome-wide analysis of deoxyadenosine methyltransferase-mediated control of gene expression in *Escherichia coli.*
Mol Microbiol. 2002 Aug;45(3):673-95.

Oshima T, Aiba H, Masuda Y, Kanaya S, Sugiura M, Wanner BL, Mori H, Mizuno T.

Transcriptome analysis of all two-component regulatory system mutants of *Escherichia coli* K-12.
Mol Microbiol. 2002 Oct;46(1):281-91.

Pascal MC, Chippaux M, Abou-Jaoude A, Blaschkowski HP, Knappe J. Mutants of *Escherichia coli* K12 with defects in anaerobic pyruvate metabolism.
J Gen Microbiol. 1981 May;124(Pt 1):35-42.

Parekh S, 2004. Strain improvement. In *The desk encyclopedia of Microbiology*. ed. Schaecter. pp. 960. London, Elsevier Academic Press.

Park MH, Wong BB, Lusk JE.
Mutants in three genes affecting transport of magnesium in *Escherichia coli*: genetics and physiology.
J Bacteriol. 1976 Jun;126(3):1096-103.

Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR.
Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7.
Nature. 2001 Jan 25;409(6819):529-33. Erratum in: Nature 2001 Mar 8;410(6825):240.

Pfeiffer J, Guhl J, Waidner B, Kist M, Bereswill S.
Magnesium uptake by CorA is essential for viability of the gastric pathogen Helicobacter pylori.
Infect Immun. 2002 Jul;70(7):3930-4.

Phadtare S, Kato I, Inouye M.

DNA microarray analysis of the expression profile of *Escherichia coli* in response to treatment with 4,5-dihydroxy-2-cyclopenten-1-one.
J Bacteriol. 2002 Dec;184(23):6725-9.

Polen T, Wendisch VF.
Genomewide expression analysis in amino acid-producing bacteria using DNA microarrays.
Appl Biochem Biotechnol. 2004 Jul-Sep;118(1-3):215-32.

Polen T, Rittmann D, Wendisch VF, Sahm H.
DNA microarray analyses of the long-term adaptive response of *Escherichia coli* to acetate and propionate.
Appl Environ Microbiol. 2003 Mar;69(3):1759-74.

Polissi A, De Laurentis W, Zangrossi S, Briani F, Longhi V, Pesole G, Deho G. Changes in *Escherichia coli* transcriptome during acclimatization at low temperature.
Res Microbiol. 2003 Oct;154(8):573-80.

Posfai G, Kolisnychenko V, Bereczki Z, Blattner FR.
Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome.
Nucleic Acids Res. 1999 Nov 15;27(22):4409-15.

Raina S, Georgopoulos C. A new *Escherichia coli* heat shock gene, htrC, whose product is essential for viability only at high temperatures.
J Bacteriol. 1990 Jun;172(6):3417-26.

Reitzer L, Schneider BL.
Metabolic context and possible physiological themes of sigma(54)-dependent genes in *Escherichia coli*.

Microbiol Mol Biol Rev. 2001 Sep;65(3):422-44, table of contents. Review.

Ren D, Bedzyk LA, Thomas SM, Ye RW, Wood TK.
Gene expression in *Escherichia coli* biofilms.
Appl Microbiol Biotechnol. 2004 May;64(4):515-24. Epub 2004 Jan 16.

Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR.
Genome-wide expression profiling in *Escherichia coli* K-12.
Nucleic Acids Res. 1999 Oct 1;27(19):3821-35.

Rode CK, Melkerson-Watson LJ, Johnson AT, Bloch CA.
Type-specific contributions to chromosome size differences in *Escherichia coli.*
Infect Immun. 1999 Jan;67(1):230-6.

Ross-Macdonald P, Sheehan A, Roeder GS, Snyder M.
A multipurpose transposon system for analyzing protein production, localization, and function in Saccharomyces cerevisiae.
Proc Natl Acad Sci U S A. 1997 Jan 7;94(1):190-5.

Rozen Y, Larossa RA, Templeton LJ, Smulski DR, Belkin S.
Gene expression analysis of the response by *Escherichia coli* to seawater.
Antonie Van Leeuwenhoek. 2002 Aug;81(1-4):15-25. Review.

Rudd, K.E., Humphery-Smith, I., Wasinger, V.C. and Bairoch, A., Low molecular weight proteins: a challenge for post-genomic research,
Electrophoresis, 19:536-544, 1998.

Russell CB, Thaler DS, Dahlquist FW.
Chromosomal transformation of *Escherichia coli* recD strains with linearized plasmids.
J Bacteriol. 1989 May;171(5):2609-13.

Salmon K, Hung SP, Mekjian K, Baldi P, Hatfield GW, Gunsalus RP.
Global gene expression profiling in *Escherichia coli* K12. The effects of oxygen availability and FNR.
J Biol Chem. 2003 Aug 8;278(32):29837-55. Epub 2003 May 15.

Sassetti CM, Boyd DH, Rubin EJ.
Comprehensive identification of conditionally essential genes in mycobacteria.
Proc Natl Acad Sci U S A. 2001 Oct 23;98(22):12712-7. Epub 2001 Oct 16.

Schembri MA, Ussery DW, Workman C, Hasman H, Klemm P.
DNA microarray analysis of fim mutations in *Escherichia coli.*
Mol Genet Genomics. 2002 Aug;267(6):721-9. Epub 2002 Jun 21.

Serres MH, Gopal S, Nahum LA, Liang P, Gaasterland T, Riley M.
A functional update of the *Escherichia coli* K-12 genome.
Genome Biol. 2001;2(9):RESEARCH0035. Epub 2001 Aug 20.

Siew N, Fischer D.
Analysis of singleton ORFans in fully sequenced microbial genomes.
Proteins. 2003 Nov 1;53(2):241-51.

Siew N, Fischer D.
Twenty thousand ORFan microbial protein families for the biologist?
Structure (Camb). 2003 Jan;11(1):7-9. Review.

Silver S, Clark D.
Magnesium transport in *Escherichia coli.*
J Biol Chem. 1971 Feb 10;246(3):569-76. No abstract available.

Schembri MA, Kjaergaard K, Klemm P.
Global gene expression in *Escherichia coli* biofilms.
Mol Microbiol. 2003 Apr;48(1):253-67.

Skovran E, Lauhon CT, Downs DM.
Lack of YggX results in chronic oxidative stress and uncovers subtle defects in Fe-S cluster metabolism in Salmonella enterica.
J Bacteriol. 2004 Nov;186(22):7626-34.

Smith RL, Maguire ME.
Microbial magnesium transport: unusual transporters searching for identity.
Mol Microbiol. 1998 Apr;28(2):217-26. Review.

Smith RL, Szegedy MA, Kucharski LM, Walker C, Wiet RM, Redpath A, Kaczmarek MT, Maguire ME.
The CorA Mg2+ transport protein of Salmonella typhimurium. Mutagenesis of conserved residues in the third membrane domain identifies a Mg2+ pore.
J Biol Chem. 1998 Oct 30;273(44):28663-9.

Snavely MD, Gravina SA, Cheung TT, Miller CG, Maguire ME.
Magnesium transport in Salmonella typhimurium. Regulation of mgtA and mgtB expression.
J Biol Chem. 1991 Jan 15;266(2):824-9.

Szent-Gyorgyi C.
A simplified method for the repeated replacement of yeast chromosomal sequences with in vitro mutations.
Yeast. 1996 Jun 15;12(7):667-72.

Tani TH, Khodursky A, Blumenthal RM, Brown PO, Matthews RG.

Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis.
Proc Natl Acad Sci U S A. 2002 Oct 15;99(21):13471-6. Epub 2002 Oct 08.

Tao H, Bausch C, Richmond C, Blattner FR, Conway T.
Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media.
J Bacteriol. 1999 Oct;181(20):6425-40.

Tong X, Campbell JW, Balazsi G, Kay KA, Wanner BL, Gerdes SY, Oltvai ZN.
Genome-scale identification of conditionally essential genes in E. coli by DNA microarrays.
Biochem Biophys Res Commun. 2004 Sep 10;322(1):347-54.

Tucker DL, Tucker N, Conway T.
Gene expression profiling of the pH response in *Escherichia coli.*
J Bacteriol. 2002 Dec;184(23):6551-8.

Wei Y, Lee JM, Richmond C, Blattner FR, Rafalski JA, LaRossa RA.
High-density microarray-mediated gene expression profiling of *Escherichia coli.*
J Bacteriol. 2001 Jan;183(2):545-56.

Wei Y, Lee JM, Smulski DR, LaRossa RA.
Global impact of sdiA amplification revealed by comprehensive gene expression profiling of *Escherichia coli.*
J Bacteriol. 2001 Apr;183(7):2265-72.

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR.

Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli.*

Proc Natl Acad Sci U S A. 2002 Dec 24;99(26):17020-4. Epub 2002 Dec 05.

Wilson RB, Davis D, Mitchell AP.

Rapid hypothesis testing with Candida albicans through gene disruption with short homology regions.

J Bacteriol. 1999 Mar;181(6):1868-74.

Winans SC, Elledge SJ, Krueger JH, Walker GC.

Site-directed insertion and deletion mutagenesis with cloned fragments in *Escherichia coli.*

J Bacteriol. 1985 Mar;161(3):1219-21.

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Davis RW, et al.

Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis.

Science. 1999 Aug 6;285(5429):901-6.

Wu CF, Valdes JJ, Bentley WE, Sekowski JW.

DNA microarray for discrimination between pathogenic 0157:H7 EDL933 and non-pathogenic *Escherichia coli* strains.

Biosens Bioelectron. 2003 Oct 30;19(1):1-8.

Woese CR.

Bacterial evolution.

Microbiol Rev. 1987 Jun;51(2):221-71.

Yamada M, Talukder AA, Nitta T. Characterization of the ssnA gene, which is involved in the decline of cell viability at the beginning of stationary phase in *Escherichia coli.* J Bacteriol. 1999 Mar;181(6):1838-46.

Yap WH, Zhang Z, Wang Y.
Distinct types of rRNA operons exist in the genome of the actinomycete Thermomonospora chromogena and evidence for horizontal transfer of an entire rRNA operon.
J Bacteriol. 1999 Sep;181(17):5201-9.

Yoshida T, Ueguchi C, Yamada H, Mizuno T. Function of the *Escherichia coli* nucleoid protein, H-NS: molecular analysis of a subset of proteins whose expression is enhanced in a hns deletion mutant.
Mol Gen Genet. 1993 Feb;237(1-2):113-22.

Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, Court DL.
An efficient recombination system for chromosome engineering in *Escherichia coli.*
Proc Natl Acad Sci U S A. 2000 May 23;97(11):5978-83.

Zheng M, Wang X, Templeton LJ, Smulski DR, LaRossa RA, Storz G.
DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide.
J Bacteriol. 2001 Aug;183(15):4562-70.

Zimmer DP, Soupene E, Lee HL, Wendisch VF, Khodursky AB, Peter BJ, Bender RA, Kustu S.
Nitrogen regulatory protein C-controlled genes of *Escherichia coli*: scavenging as a defense against nitrogen limitation.
Proc Natl Acad Sci U S A. 2000 Dec 19;97(26):14674-9.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M, Snyder M.
Global analysis of protein activities using proteome chips.
Science. 2001 Sep 14;293(5537):2101-5. Epub 2001 Jul 26.

Zhu H, Bilgin M, Snyder M.
Proteomics.
Annu Rev Biochem. 2003;72:783-812. Review.

Appendix 1											
Strain	LB 37	LB 30	LB 45	LB an	M9 37	M9 30	pH5.8	pH5.6	pH9	pH9.2	Co 1mM
MG1655	9E+08	2E+08	1E+09	9E+08	1E+09	9E+08	6E+08	9E+08	2E+09	1E+04	7E+08
ycfR	4E+08	8E+08	4E+08	1E+09	5E+08	5E+08	6E+08	6E+08	1E+09	1E+04	2E+08
yceP	2E+09	1E+09	2E+09	2E+09	9E+08	7E+08	2E+09	1E+09	2E+09	1E+04	5E+08
htrC	8E+08	1E+09	8E+08	5E+08	1E+09	8E+08	7E+08	7E+08	9E+08	1E+04	3E+08
yhcN	7E+08	8E+08	7E+08	6E+08	3E+08	6E+08	3E+08	8E+08	8E+08	1E+04	5E+08
yahO	8E+08	4E+08	1E+09	9E+08	5E+08	4E+08	5E+08	5E+08	3E+08	1E+04	9E+08
ybiM	8E+08	5E+08	1E+09	4E+08	5E+08	6E+08	9E+08	6E+08	5E+08	1E+04	3E+08
ybiJ	8E+08	5E+08	1E+09	4E+08	5E+08	6E+08	9E+08	6E+08	5E+08	1E+04	3E+08
ydgH	6E+08	6E+08	9E+08	9E+08	5E+08	4E+08	4E+08	5E+08	5E+08	2E+04	7E+08
yjfY	6E+08	2E+08	5E+08	3E+08	9E+08	6E+08	1E+08	3E+08	6E+08	1E+04	2E+08
ykgI	2E+09	3E+09	2E+09	2E+09	1E+09	1E+09	1E+09	2E+09	2E+09	1E+04	2E+09
ybhC	5E+08	5E+08	7E+08	2E+08	9E+08	3E+08	6E+08	6E+08	4E+08	1E+04	6E+08
ycjT	7E+08	6E+08	7E+08	1E+09	3E+08	6E+08	8E+08	2E+08	7E+08	1E+04	5E+08
ydeK	3E+08	8E+08	6E+08	3E+08	4E+08	2E+08	5E+08	4E+08	7E+08	1E+04	5E+08
yedJ	5E+08	4E+08	3E+08	3E+08	3E+08	4E+08	3E+08	3E+08	4E+08	2E+04	2E+08
yegI	6E+08	7E+08	9E+08	1E+09	4E+08	2E+08	5E+08	4E+08	6E+08	1E+04	6E+08
yegR	7E+08	5E+08	7E+08	8E+08	4E+08	5E+08	6E+08	4E+08	7E+08	2E+04	4E+08
yeiN	6E+08	8E+08	8E+08	5E+08	2E+08	7E+08	3E+08	5E+08	4E+08	1E+04	3E+08
yfbL	9E+08	1E+09	7E+08	1E+09	5E+08	7E+08	5E+08	2E+08	4E+08	2E+04	5E+08
yfjP	1E+09	4E+08	1E+09	6E+08	6E+08	4E+08	4E+08	6E+08	9E+08	1E+04	6E+08
ygaQ	4E+08	6E+08	6E+08	8E+08	6E+08	7E+08	6E+08	8E+08	3E+08	1E+04	4E+08
ygiN	7E+08	5E+08	1E+09	1E+09	6E+08	3E+08	3E+08	9E+08	6E+08	2E+04	4E+08
ygjMN	8E+08	4E+08	6E+08	1E+09	5E+08	1E+09	1E+09	5E+08	1E+09	2E+04	5E+08
yhiM	1E+09	9E+08	6E+08	8E+08	4E+08	1E+09	8E+08	7E+08	7E+08	1E+04	5E+08
yigE	5E+08	9E+08	7E+08	6E+08	6E+08	2E+09	7E+08	8E+08	4E+08	2E+04	7E+08
yihR	6E+08	9E+08	1E+09	5E+08	5E+08	4E+08	3E+08	9E+08	6E+08	2E+04	1E+08
yjdA	3E+08	7E+08	3E+08	3E+08	4E+08	3E+08	6E+08	7E+08	9E+08	2E+04	7E+08
yjiW	2E+09	2E+09	2E+09	3E+09	6E+08	1E+09	2E+09	3E+09	2E+09	2E+04	7E+08
yncE	7E+08	1E+09	6E+08	1E+09	6E+08	1E+09	4E+08	8E+08	1E+09	1E+04	3E+08
ypjC	9E+08	8E+08	1E+09	6E+08	7E+08	6E+08	7E+08	9E+08	8E+08	1E+04	1E+08
yqhG	7E+08	1E+09	9E+08	8E+08	2E+08	8E+08	5E+08	3E+08	5E+08	2E+04	1E+08
yraQ	7E+08	7E+08	1E+09	1E+09	3E+08	7E+08	4E+08	6E+08	1E+09	1E+04	5E+08
ydhR-Z	5E+08	5E+08	6E+08	1E+08	6E+08	2E+08	5E+08	2E+08	2E+08	1E+04	5E+08
ydhV	5E+08	3E+08	3E+08	8E+08	5E+08	2E+08	5E+08	5E+08	7E+08	2E+04	1E+09
hdeA	2E+08	7E+08	8E+08	4E+08	8E+08	4E+08	7E+08	9E+08	8E+08	1E+04	8E+08
hdeB	2E+08	7E+08	8E+08	4E+08	8E+08	4E+08	7E+08	9E+08	8E+08	2E+04	8E+08
yahLM	7E+08	5E+08	8E+08	7E+08	4E+08	4E+08	4E+08	7E+08	4E+08	1E+04	3E+08
yjdI-K	2E+08	5E+08	6E+08	8E+08	8E+08	2E+08	4E+08	6E+08	2E+09	1E+04	2E+08
day test stopp	2	3	3	3	3	4	3	4	2		3

Co 1.5mM	Cu 5mM	Cu 6mM	Ni 2mM	Ni 3mM	Zn 2mM	Zn 3mM	NaCl 1.2mM	NaCl .8mM	CV 10ug/ml	CV 20ug/ml	Strain
2E+07	2E+09	5E+07	6E+08	6E+08	8E+07	1E+07	8E+08	7E+08	3E+08	3E+07	MG1655
5E+07	1E+09	3E+08	5E+08	5E+08	1E+07	1E+07	4E+08	6E+08	1E+08	2E+07	*ycfR*
2E+07	2E+09	1E+08	1E+09	6E+08	2E+08	2E+07	1E+08	6E+08	2E+08	6E+06	*yceP*
3E+07	2E+09	2E+08	2E+08	2E+08	3E+08	1E+07	2E+08	9E+08	1E+08	3E+07	*htrC*
3E+07	1E+09	3E+08	5E+08	5E+08	2E+08	1E+07	5E+08	4E+08	2E+08	2E+07	*yhcN*
4E+07	1E+09	2E+08	6E+08	4E+08	4E+08	1E+07	5E+08	7E+08	3E+08	1E+07	*yahO*
1E+07	4E+08	5E+08	1E+08	1E+08	4E+08	3E+07	2E+08	1E+09	2E+08	3E+07	*ybiM*
2E+07	8E+08	7E+08	6E+08	2E+08	1E+08	2E+07	2E+08	5E+08	1E+08	1E+07	*ybiJ*
2E+07	2E+09	5E+07	5E+08	2E+08	3E+08	1E+07	1E+08	2E+08	2E+08	1E+07	*ydgH*
2E+07	6E+08	6E+07	3E+08	1E+08	2E+08	1E+07	1E+08	8E+08	1E+08	2E+07	*yjfY*
1E+07	1E+09	2E+08	2E+09	1E+09	1E+08	3E+06	4E+08	6E+08	3E+08	1E+07	*ykgI*
3E+07	1E+09	3E+08	4E+08	3E+08	1E+08	1E+07	5E+08	1E+09	1E+08	3E+07	*ybhC*
3E+07	1E+09	7E+08	5E+08	3E+08	3E+08	1E+07	3E+08	1E+09	3E+08	4E+07	*ycjT*
1E+07	1E+09	1E+08	4E+08	3E+08	4E+07	2E+07	1E+08	6E+08	4E+08	3E+06	*ydeK*
1E+07	1E+09	5E+08	3E+08	2E+08	2E+08	1E+07	6E+08	4E+08	1E+08	2E+07	*yedJ*
7E+07	1E+09	8E+08	3E+08	2E+08	2E+08	3E+07	3E+08	1E+09	3E+08	3E+07	*yegI*
2E+07	2E+09	6E+08	6E+08	6E+08	1E+08	1E+07	3E+08	8E+08	2E+08	2E+07	*yegR*
2E+07	2E+09	1E+08	5E+08	5E+08	1E+08	2E+07	1E+08	9E+08	2E+08	3E+07	*yeiN*
2E+07	1E+09	9E+08	4E+08	1E+08	1E+08	2E+07	2E+08	8E+08	4E+08	1E+07	*yfbL*
5E+07	1E+09	1E+08	9E+08	3E+08	1E+08	2E+07	1E+08	1E+08	3E+08	2E+07	*yfjP*
2E+07	9E+08	3E+08	2E+08	7E+08	2E+08	2E+07	2E+08	2E+08	3E+08	1E+07	*ygaQ*
2E+07	2E+09	5E+08	6E+08	1E+08	2E+08	4E+07	3E+08	2E+09	1E+08	1E+07	*ygiN*
3E+07	1E+09	1E+08	5E+08	5E+08	1E+08	3E+07	4E+08	7E+08	1E+08	5E+06	*ygjMN*
2E+07	7E+08	2E+08	5E+08	4E+08	6E+07	1E+07	2E+08	9E+08	2E+08	2E+07	*yhiM*
2E+05	1E+09	5E+08	6E+08	4E+08	2E+08	7E+07	4E+08	1E+09	4E+08	3E+07	*yigE*
1E+07	3E+08	3E+08	2E+08	3E+08	6E+07	3E+07	2E+08	6E+08	3E+08	2E+07	*yihR*
1E+07	1E+09	1E+08	9E+08	5E+08	9E+07	3E+07	4E+08	9E+08	1E+08	4E+07	*yjdA*
2E+07	6E+08	3E+08	1E+09	8E+08	2E+08	2E+07	6E+08	1E+09	3E+08	4E+07	*yjiW*
6E+07	1E+09	2E+08	7E+08	5E+08	2E+08	6E+06	4E+08	7E+08	2E+08	3E+07	*yncE*
3E+07	5E+08	5E+07	4E+08	5E+08	2E+08	4E+06	3E+08	1E+09	1E+08	2E+07	*ypjC*
1E+07	2E+09	5E+07	5E+08	2E+08	2E+08	2E+07	2E+08	1E+09	5E+08	1E+07	*yqhG*
1E+07	9E+08	1E+08	4E+08	1E+08	1E+08	3E+07	4E+08	6E+08	3E+08	3E+07	*yraQ*
1E+07	9E+08	4E+07	4E+08	3E+08	3E+08	1E+07	1E+08	9E+08	4E+08	2E+07	*ydhR-Z*
4E+07	2E+09	3E+08	3E+08	2E+08	4E+08	1E+07	2E+08	8E+08	2E+08	4E+07	*ydhV*
5E+07	1E+09	7E+08	7E+08	2E+08	2E+08	2E+07	3E+08	6E+08	2E+08	2E+07	*hdeA*
5E+07	6E+08	7E+07	1E+08	2E+08	1E+08	1E+07	1E+08	1E+09	1E+08	1E+07	*hdeB*
2E+07	7E+08	3E+08	3E+08	1E+09	3E+08	1E+07	3E+08	7E+08	4E+08	2E+07	*yahLM*
3E+07	1E+09	1E+08	1E+08	4E+08	1E+08	4E+06	3E+08	8E+08	3E+08	2E+07	*yjdI-K*
6	2	3	3	4	3	5	4	2	2	4	